



# A Survey on Privacy-Preserving Caching at Network Edge: Classification, Solutions, and Challenges

**XIANZHI ZHANG**, School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China and School of Computing, Macquarie University, Sydney, Australia

**YIPENG ZHOU**, School of Computing, Macquarie University, Sydney, Australia

**DI WU**, School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China

**QUAN Z. SHENG**, School of Computing, Macquarie University, Sydney, Australia

**SHAZIA RIAZ**, School of Computing, Macquarie University, Sydney, Australia and Department of Computer Science, Government College Women University Faisalabad. Pakistan, Faisalabad, Pakistan

**MIAO HU**, School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China

**LINCHANG XIAO**, School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China

Caching content at the edge network is a popular and effective technique widely deployed to alleviate the burden of network backhaul, shorten service delay, and improve service quality. However, there has been some controversy over privacy violations in caching content at the edge network. On the one hand, the multi-access open edge network provides an ideal entrance or interface for external attackers to obtain private data from edge caches by extracting sensitive information. On the other hand, privacy can be infringed on by curious edge caching providers through caching trace analysis targeting the achievement of better caching performance or higher profits. Therefore, an in-depth understanding of privacy issues in edge caching networks is vital and indispensable for creating a privacy-preserving caching service at the edge network. In this article, we are among the first to fill this gap by examining privacy-preserving techniques for caching content at the edge network. First, we provide an introduction to the background of privacy-preserving edge caching. Next, we summarize the key privacy issues and present a taxonomy for caching at the edge network from

---

This work was supported in part by the Shenzhen Science and Technology Program under grant KJZD20230923113901004, in part by the Science and Technology Planning Project of Guangdong Province under grant 2023A0505020006, in part by the Science and Technology Development Fund, Macau SAR, under grant 0008/2022/AGJ, and in part by the National Natural Science Foundation of China under grants U2001209 and 62072486.

Authors' Contact Information: Xianzhi Zhang, School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, Guangdong, China and School of Computing, Macquarie University, Sydney, New South Wales, Australia; e-mail: zhangxzh9@mail2.sysu.edu.cn; Yipeng Zhou, School of Computing, Macquarie University, Sydney, New South Wales, Australia; e-mail: yipeng.zhou@mq.edu.au; Di Wu (Corresponding author), School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, Guangdong, China; e-mail: wudi27@mail.sysu.edu.cn; Quan Z. Sheng, School of Computing, Macquarie University, Sydney, New South Wales, Australia; e-mail: michael.sheng@mq.edu.au; Shazia Riaz, School of Computing, Macquarie University, Sydney, New South Wales, Australia and Department of Computer Science, Government College Women University Faisalabad. Pakistan, Faisalabad, Punjab, Pakistan; e-mail: Shazia.Riaz@mq.edu.au; Miao Hu, School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, Guangdong, China; e-mail: humiao@outlook.com; Linchang Xiao, School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, Guangdong, China; e-mail: xiaolch3@mail2.sysu.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 0360-0300/2025/01-ART114

<https://doi.org/10.1145/3706630>

the perspective of private information. Additionally, we conduct a retrospective review of the state-of-the-art countermeasures against privacy leakage from content caching at the edge network. Finally, we conclude the survey and envision challenges for future research.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Security and privacy** → **Privacy protections**; • **Networks** → **Network privacy and anonymity**; • **Information systems** → **Distributed storage**; • **Computer systems organization** → **Distributed architectures**;

Additional Key Words and Phrases: Edge cache, privacy-preserving caching, edge networks, countermeasure, caching performance

**ACM Reference Format:**

Xianzhi Zhang, Yipeng Zhou, Di Wu, Quan Z. Sheng, Shazia Riaz, Miao Hu, and Linchang Xiao. 2025. A Survey on Privacy-Preserving Caching at Network Edge: Classification, Solutions, and Challenges. *ACM Comput. Surv.* 57, 5, Article 114 (January 2025), 38 pages. <https://doi.org/10.1145/3706630>

---

## 1 Introduction

Content caching at the edge network is driven by two factors. First, the population of networked devices has become astronomical due to advances in intelligent terminals and the broad deployment of the **Internet of Things (IoT)** [15, 16, 26, 118]. Second, the Internet content market is blooming due to the proliferation of various multimedia content [54, 114]. According to a report by Splunk, a Cisco company, there will be approximately 5.44 billion Internet users worldwide in 2024, including 5.07 billion social media users [72]. As a result, network-based content delivery services are extremely bandwidth consuming. At the same time, emerging network technologies, such as Gigabit Ethernet, and 5G and beyond, are expected to provide extremely high data transmission rates and low access delays for terminal devices at the edge network to support time-sensitive services such as autonomous driving, industrial automation, high-quality video streaming, and virtual/enhanced emerging applications.

Such a vast dataflow brings two main challenges to the established networks. First, it brings a heavy communication burden to the Internet core network links. During the peak hours of network usage, a large amount of content transmission will inevitably aggravate the link burden of the core network, causing network congestion and increasing network operating costs. Second, it will also prolong the service delay of content transmission from remote servers to end devices, which will adversely influence users' service **Quality-of-Experience (QoE)** or even ruin the reliability of delay-sensitive applications.

Edge caching is a technique that involves storing content in close proximity to end users, typically at or near the point of user access or ahead of the core network [54]. Its primary objective is to shorten service latency and enhance content delivery performance by bringing content closer to the users who request it. When users request content that is available in **edge caches (ECs)**, their requests can be directly served at the edge network with a high **Quality-of-Service (QoS)**. However, if the requested content is not available in the EC, it can be redirected to a remote server, such as a data center. Here, we make a brief introduction to edge caching from five aspects.

**Benefit of Edge Caching.** Caching content at the edge network is effective in reducing the burden of network backhaul [33, 62, 98], shortening service latency [17, 62, 107], and diminishing resource cost [28, 33]. First, it is common to cache popular content at the edge network through which the edge network can offload the access of requests and hence reduce the backhaul dataflow. Even though the caching capability is limited at the edge network, ECs can offload up to 35% of the traffic burden over backhaul links [54]. Second, the service latency can be shortened by caching content on edge devices near end users. In particular, a shortened latency is critical for content delivery of latency-sensitive applications [54]. Third, edge networks can make content

access inexpensive since caching content at edge devices can avoid the bottleneck. For example, in wireless edge networks, spectral efficiency and energy efficiency can be improved by about 900% and 500%, respectively, by using edge networks for caching content [43].

**Where to Cache.** Building on the work of Ni et al. [54], we further identify three main entities in edge networks for edge caching as follows. First, *end devices* (e.g., smartphones, laptops, intelligent vehicles, and Industrial IoT devices) carried by users will generate requests for downloading content via networks [16, 26]. It is possible that end devices can share content through D2D (device-to-device) communications with licensed-band or unlicensed-band protocols. Second, *access infrastructures*, utilizing wired and/or wireless communication technologies, can support end devices in accessing the Internet. These infrastructures include 5G **small base stations (SBSs)** [94], WiFi routers, local switches, and **roadside units (RSUs)** in the **Internet of Vehicles (IoV)** [14, 113]. Popular content can be cached within these access infrastructures to promptly serve user requests. Third, **edge servers (ESs)** positioned ahead of the core network, such as **edge nodes (ENs)** in the **Content Delivery Network (CDN)** [17], edge routers in the **Information-Centric Network (ICN)** [71, 95, 96], and macro base stations operated by **Internet service providers (ISPs)** [7], can be utilized as ECs, a concept known as *in-network edge caching*. These ESs, typically maintained by third-party suppliers, are the core points for multi-access edge networks, enhancing various content delivery applications.

**What to Cache.** In edge caching systems, determining what content to cache is crucial for optimizing cache space utilization and reducing latency. The content to be cached generally falls into three categories. The category of *user-related popular content* includes content that is frequently requested by end users, such as web pages, videos, images, and other multimedia files [17, 85]. Caching such content at the edge improves user experience by reducing service delay when users access commonly accessed content. The category of *public and static content* includes high-reuse, non-user-specific content associated with applications [54], such as JavaScript files, CSS stylesheets, icons, PDF documents, and API responses. Caching these static resources decreases application load times and reduces the burden on central servers. The category of *edge-computable and storable content* includes data that can be computed and stored directly at the edge, such as model parameters for **federated learning (FL)** [45, 62], user patterns and content popularity for edge caching decisions [14, 17], and IoT sensor data awaiting processing [84, 101]. Caching such content helps minimize backhaul traffic, enables efficient edge processing, and reduces overall latency by avoiding redundant computations.

**How to Cache.** Edge caching strategies can be broadly classified into reactive and proactive approaches. The *reactive caching* approach employs eviction-based methods that decide whether to cache a specific content item only after it has been requested. This approach often relies on empirical formulas and classical caching algorithms, such as LRU (Least Recently Used) and LFU (Least Frequently Used), as well as their variants [21, 68]. While these algorithms are simple and efficient, they frequently encounter challenges in selecting optimal parameter values, which can limit their performance in dynamic and diverse edge environments. The *proactive caching* approach involves predictive methods to determine what content should be cached before any user requests are made. This approach leverages content popularity predictions and user behavior profiles to make caching decisions at edge networks in advance. Advanced machine learning models, such as LSTM (Long Short-Term Memory) networks [22], are often employed to forecast content demand based on historical request patterns in proactive caching. These learning-driven methods generally offer superior caching performance compared to classical algorithms by automatically adjusting model parameters. However, they may require extensive computational capacity and high-quality training data, both of which may be often limited at the edge.

**Privacy Concerns of Edge Caching.** Despite the enormous benefits brought by caching content at the edge network, there has been some controversy over privacy violations brought by such caching. The concerns can be illustrated from two aspects. The first privacy threat comes from external attackers, such as malicious user devices [2, 17, 42, 61, 71, 75]. The multi-access open property of the edge network provides an ideal entrance or interface for external attackers to obtain the cached content from the EC to extract sensitive information of end users [93]. Adversaries can obtain user-sensitive information by launching cache side-channel attacks [2, 42, 71] and cache tampering attacks [17, 61, 75]. However, it is non-trivial to embed advanced privacy protection mechanisms into edge networks due to the limited computing capacity, energy power, and storage space of edge devices. Second, user privacy can be infringed by curious edge caching providers by analyzing traces and management records. Due to limited caching space relative to the rapidly growing user population and the scale of content [122], edge network providers have a strong motivation to spy on user privacy to improve their resource utilization. In other words, if content popularity can be accurately predicted, the right content can be cached by edge devices just before the surge of requests toward content [114]. Hence, edge network providers are curious about users' personal interests and confidential information to infer their request behaviors, which can be extracted from users' historical request traces (e.g., request patterns [14, 17, 114], identifiable information [7, 14, 17, 113, 124]). Edge network providers can implement monitoring attacks [96, 113] and inference attacks [45, 62] in their systems to compromise users' privacy based on collected request information from users. Therefore, an in-depth understanding of privacy risks in **privacy-preserving edge caching (PPEC)** is crucial for the design of feasible solutions to achieve a privacy-preserving content cache at the edge network.

**Our Contributions.** Recently, significant progress has been made in enhancing privacy protection for content caching at edge networks. However, these works fail to provide a comprehensive discussion of the privacy issues in edge caching systems. For instance, Ren et al. [63] primarily discuss the state-of-the-art research on caching and privacy, respectively, in emerging edge computing paradigms. Besides, most surveys only discuss privacy issues for particular scenarios, such as IoT [23, 65], edge intelligence [108], the metaverse [57], and FL [8], while overlooking the distinct aspects of edge caching. Other surveys [54, 87] addressed privacy-preserving solutions and countermeasures for edge caching without covering all relevant issues in a thorough manner. In particular, there has been a lack of comprehensive discussion of protection methods targeting different types of private information in PPEC. Given these limitations and the absence of comprehensive literature reviews, this article aims to thoroughly examine and categorize current works on privacy issues in edge caching scenarios. The main contributions of this article are summarized as follows. First, we present in-depth discussions on sensitive information in edge caching and propose a taxonomy from a private information perspective to classify existing works. To the best of our knowledge, this is the first such comprehensive exposition. Second, we conduct a thorough review of recent high-quality research, diving into the background of privacy attacks and mitigation methods in the realm of edge caching. Our review encompasses the latest solutions proposed for enhancing privacy in edge caching, which have been published in leading conferences and journals in the fields of computing networks, architecture, and privacy, such as CCS, INFOCOM, ToN, JSAC, TPDS, TIFS, and TDSC, as well as other top venues. Based on different kinds of privacy information and attacks toward each kind of privacy information, we respectively review countermeasures to defend against attacks for protecting each kind of infringed privacy. Third, based on open problems outlined in existing works, we envision privacy-related open challenges in PPEC to provide insights for inspiring future research.

**Article Outline.** The rest of this article is organized as follows. Section 2 provides an introduction to the taxonomy of privacy-preserving solutions that are based on the protection of sensitive

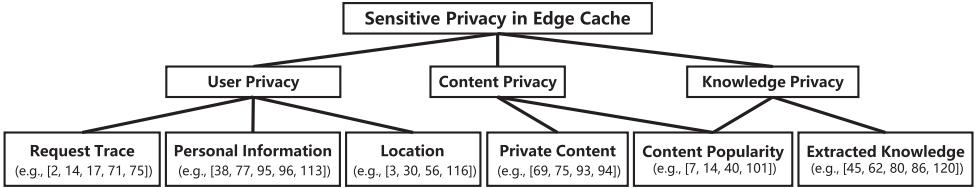


Fig. 1. The framework for PPEC encompasses six distinct types of data concerns, which can be primarily classified into three categories of private information: user privacy, content privacy, and knowledge privacy.

information data in edge caching. Section 3 provides a background discussion on privacy issues in the edge caching paradigm from two plain perspectives: privacy attacks and mitigation methods. From Section 4 to Section 6, we describe the possible privacy mitigation solutions for edge caching in correspondence with three main classes of privacy: user privacy, content privacy, and knowledge privacy, respectively. Section 7 provides open challenges and future research directions. Finally, we present a summary in Section 8. To facilitate readability, we have compiled a summary of commonly used abbreviations for the solutions in Table 1 in the appendix.

## 2 Overview of Private Information in Edge Caching

In this section, we overview sensitive information that should be protected to avoid privacy leakage in PPEC. In the realm of edge caching, sensitive information can be exposed by either users [2] unconsciously or ESs [7, 14]. Specifically, users' sensitive information includes personal information, browsing history, location, and private content data, through their request traces to the ES or other service providers. Similarly, ESs can leak their private information and extract knowledge from a collection of users who have interacted with ESs [14, 17]. Therefore, to build a privacy-preserving content caching system, the first step is to understand what private information can be exposed by users and ESs. In Figure 1, we outline all kinds of sensitive information that should be protected in PPEC. We will elaborate on each kind of private information in this section.

### 2.1 User Privacy

In PPEC, all information related to users but not directly related to cached content is regarded as user privacy such as users' historical records, age, gender, and location. For our discussion, we classify all user privacy information into three types: request trace, personal information, and location.

**2.1.1 Request Trace.** A *request trace* refers to a sequence of content requests and responses between an end device and ESs or service providers. These traces often contain private information such as request patterns [2, 42], preferences [14, 61], and interests [17, 71]. Advertisers or malicious attackers can exploit such information to make profits or harm. Additionally, user request traces are valuable assets to service providers and caching systems. Service providers or **content providers (CPs)** can analyze these request traces to infer users' behavior patterns, such as the type of websites or applications they frequently use and the content they prefer to consume. ESs can maintain and analyze request traces to improve caching performance by predicting future requests, allowing for prefetching and caching popular content in advance.

There are primary two risks associated with request traces: interception and misuse. First, request records can be *intercepted* and sniffed by other users and external attackers. For example, malicious users can use timing attacks [2, 71] to impersonate legitimate users, sending requests to the server. Attackers may then infer user request traces by exploiting the timing difference between cached and non-cached responses [2, 71], facilitating illegal advertising and cache

pollution attacks [85]. Second, ESs and service providers, curious about user interest patterns, may *misuse* request traces for their purposes. For example, request traces can be exploited to develop trace-driven content caching algorithms, posing privacy threats from untrusted or profit-driven third-party ESs [7, 14, 66, 75].

However, designing methods to preserve user privacy in edge caching systems is non-trivial. Most existing privacy-enhancing approaches fail to effectively address the privacy leakage risks that users face in caching systems, as request records cannot be arbitrarily altered or obfuscated by users and must remain visible to service providers and edge caching servers to provide reliable services.

**2.1.2 Personal Information.** *Personal information* is a type of private information that can be mined to identify a specific end device or user in the network. Edge caching servers and service providers can obtain various types of personal information from users, depending on the specific context and implementation of the edge caching system. Typical examples of personal information that can be compromised in edge caching include the following. The first is identifier information such as pseudonyms and IP addresses. In particular, through IP addresses, we can identify a user's ISP, approximate location, and other information, with which the EC can carry out sensitive operations, such as integrity verification [75] and cache admission control [95, 96]. The second is device information such as the operating system, connection type, browser type, and version, which is also essential for ESs to provide high-performance edge caching and tailored content to users [16, 113]. The third is account-related information such as email address, gender, age, payment, and social relation, which can be captured by ECs or service providers when a user logs in or creates an account to access the service, potentially revealing more personal privacy [17, 113].

Excessively exposing personal information by edge caching can result in annoying tracking and profiling. When personal information is collected, edge caching servers and service providers can create detailed user profiles, encompassing browsing habits and interests. By identifying specific users or user groups, service providers can accurately predict future requests, allowing for content prefetching to reduce latency and improve QoS. Additionally, detailed profiles facilitate targeted advertising and personalized recommendations, potentially increasing revenue. However, these practices raise ethical concerns [16, 113], including the potential for manipulation or discrimination against certain user groups. In addition, malicious nodes and attackers can take advantage of excessive disclosure of personal information to gain unauthorized access to user accounts [95, 96] and pull off cache tampering attacks [16, 17, 75], resulting in financial losses and other harms.

**2.1.3 Location.** *Location* information is a critical type of privacy data carrying location, spatial coordinates, and the current time of moving objects. In edge caching systems, there are two fundamental types of location information: users' location information and **Points of Interest (POIs)**. When users access edge caching systems, they may unconsciously expose private location information in the following processes. First, a user's geographic location can be exposed to the EC when accessing content or services directly from the EC [18]. Second, CPs and edge caching providers can proactively collect users' geographic location information to provide better content distribution services, such as predicting user moving patterns [116]. Third, in **location-based services (LBSs)**, users may provide their private geographic information and POIs to search for their interests in the EC [3, 18, 25, 56]. This information can be abused, resulting in undesired tracking and profiling or even more severe consequences, such as location-based attacks.

Location information is sensitive and can be utilized to learn an individual's daily routine and movements. Service providers can use this information to deliver more relevant advertisements and cached content to users, potentially boosting profits. Yet, if malicious attackers obtain location information, it can put users at risk of physical harm. Malicious attackers can use location

information to track a user's movement trajectories and potentially cause harm, particularly in the case of stalking or other criminal activities.

## 2.2 Content Privacy

Content privacy refers to privacy information contained by the content stored and transmitted through edge caching systems, mainly including private content and content popularity.

**2.2.1 Private Content.** *Private content* refers to sensitive and confidential data that is stored and potentially cached by edge systems. We name such sensitive content data in edge caching systems as *private content*. Given its sensitive nature, private content requires strict privacy protections to prevent unauthorized access and misuse. This type of content includes but is not limited to video clips, photos, social media, and textual data from users, copyrighted materials, confidential business documents, and government secrets. For example, in mobile social networks, each user can be regarded as a CP who can produce fresh content desiring that their content can be efficiently and accurately delivered to consumers [94, 122]. In this case, edge computing is a feasible architecture for caching and delivering the content. Consumers in proximity [93, 116] or with close social relations [81] to a particular user CP in social networks are more likely to request this content. Thereby, using an ES to cache and deliver content in mobile social networks can diminish bandwidth costs, which however raises privacy leakage risks.

Briefly speaking, private content privacy can be infringed in several ways. First, ESs are not trustworthy and can expose cached content to the public. Second, malicious and unauthorized users at the edge network can access cached content during transmission or processing between end users and the EC or between different ECs. For instance, in cache side-channel attacks [42, 71], attackers attempt to access cached content by sending targeted requests, potentially allowing them to view sensitive information. For another instance, attackers can lodge cache tampering by injecting malicious content into the cache to exploit vulnerabilities in end user systems or steal sensitive information [17, 61, 75].

**2.2.2 Content Popularity.** *Content popularity* can be defined as the relative frequency of a particular content to be requested by users. It indicates the level of popularity of content among users. The popularity information is broadly utilized in improving caching efficiency, and caching the most popular content can effectively lower the content delivery cost. However, the popularity information is sensitive, unveiling the private preference information of users [17, 101]. Besides, it is possible that content popularity information can reveal sensitive information about CPs, such as their financial success and strategic direction, which should be kept confidential [7, 17].

The popularity information is crucial for making effective edge caching decisions but is highly susceptible to leakage. First, popularity information may be leaked during cooperative caching decision-making processes among ECs. For instance, as the number of records owned by a single ES is limited, CPs may need to provide supplementary information [7]. Additionally, edge caching servers may exchange popularity information to optimize caching decisions across the entire system [17, 101]. Privacy leakage can occur because ESs might be untrusted, or the edge environment itself may be vulnerable to attacks [17]. Second, popularity information can also be compromised through well-decided caching content. For example, through broadcasting cached content lists [14, 54] or timing attacks [71, 113], malicious entities can infer which content is more popular. This sensitive information, once exposed and tampered with, can be exploited to obtain illegal benefits, manipulate cache performance [17, 75], or even launch cache tampering attacks [17, 93]. Especially, as content popularity can describe specific content attributes and serve as key knowledge to improve caching efficiency, we consider it a unique type of information that intersects both content privacy and knowledge privacy, as shown in Figure 1.

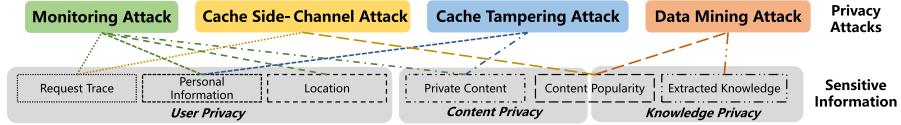


Fig. 2. The possible privacy attacks on different sensitive information in edge caching systems.

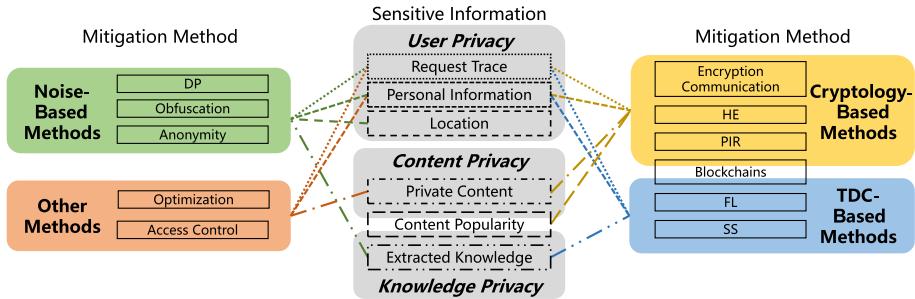


Fig. 3. The corresponding defense methods for enhancing different private information in edge caching systems.

### 2.3 Knowledge Privacy

*Knowledge privacy* refers to the insights, patterns, and parameters derived from datasets processed by machine learning models, typically owned by ECs or other service providers. Unlike user privacy and content privacy, which focus on data directly linked to users or content, knowledge privacy involves higher-level abstractions extracted from these data sources. In edge caching, service providers are particularly interested in the knowledge extracted from original datasets, as it is valuable for improving caching performance. For example, by leveraging prediction models based on this extracted knowledge, providers can make effective caching decisions in dynamic scenarios [51, 98], leading to significant improvements in edge caching performance [20, 47, 113, 114, 116]. Learning-based methods offer a feasible framework for making effective edge caching decisions, but they also pose risks of private information leakage during model training and prediction phases. Therefore, it is crucial to carefully consider and mitigate these privacy risks when employing learning-based methods for edge caching.

## 3 Overview of Attack and Defense Methods

This section is divided into two parts: an overview of attack methods targeting each type of sensitive information in edge caching systems and a summary of defense methods against each type of attack. In Figures 2 and 3, we present a relational map that illustrates the connections between potential privacy attacks, defense methods, and sensitive information in edge caching systems. In the rest of this section, we briefly discuss each type of attack and defense methods as depicted in Figure 2 and Figure 3, respectively.

### 3.1 Privacy Attack in Edge Caching Systems

There are mainly four types of privacy attacks in edge caching systems, which are monitoring attacks, data mining attacks, cache side-channel attacks, and cache tampering attacks. We introduce these attacks with potential risk entities in this subsection.

**3.1.1 Monitoring Attack.** *Monitoring attacks*, also known as eavesdropping attacks, can be divided into two main categories. The first is sniffing attacks on network communications—that is,

an adversary sniffs on network traffic through the edge caching node to read or intercept private information in network packets [113]. For example, the EC can monitor user requests during the caching service process. In other words, the edge caching operator can monitor users' requests intended to responding end users' requests and improve the caching efficiency. Through subsequent data analysis, edge caching managers can improve the caching efficiency and reduce the transmission delay of the requested content. However, a user request may contain private information, such as personal content preference [14, 66, 104], location [116], content popularity [14], and other personal information [34, 95, 96]. Therefore, edge caching systems should take both caching efficiency and privacy preservation into account. Entities that can implement sniffing attacks in network communications include edge caching managers (e.g., CPs [14], location service providers [116], ISPs or a base station [104], edge devices [14, 66, 75, 94, 122]), malicious end devices [14, 55, 95, 96], and external adversaries [113]. The second type of monitoring attack is supervisory attacks on cached content—that is, attackers conduct improper monitoring, replacement, pollution, and other privacy attack activities on cached content. By leveraging the illegal cache access, adversaries can obtain private data or information such as content popularity [4, 7, 17], user preferences [61], and other private information [17, 75]. If the cached content is not protected prudently, the user's privacy can be seriously compromised by ECs, which are often deployed by honest but curious third parties (e.g., ISPs [4, 7], ESs [17, 61, 75], and end devices [17, 61]).

**3.1.2 Data Mining Attacks.** *Data mining attacks* usually occur when an edge caching entity applies a learning-based caching algorithm to explore sensitive data for making caching decisions. Due to the high dynamics and complicated access patterns driven by users' interest [51, 98], designing an intelligent edge caching algorithm is essential to improve the caching performance. Commonly, learning-based methods make caching decisions by exploiting historical information to train a prediction model. It is necessary to feed the model training with private and sensitive data related to users, and thus users may be reluctant to share. Since edge caching decisions are generated by learning algorithms, edge caching becomes a tradeoff problem between caching performance and privacy protection level. As a consequence, learning-based methods in edge computing assisted caching are usually vulnerable to two types of privacy risks: (1) *exploratory*, in which adversaries investigate vulnerabilities (e.g., the training dataset, model parameters, and gradient data) without changing the training process, and (2) *causative*, in which attackers manipulate and inject misleading training datasets to alter the machine learning model's training process [76]. Additionally, previous research has shown that model parameters [70] and gradients [1, 117] of the machine learning model can be utilized to recover original sensitive and private information. Learning-based methods provide a practical framework for making edge caching decisions but are susceptible to privacy risks that can compromise user privacy. The potential adversaries to launch data mining attacks include edge caching managers (e.g., CPs [14, 62], ISPs [84], edge devices [45, 62, 80]) and malicious end devices [81].

**3.1.3 Cache Side-Channel Attacks.** In *cache side-channel attacks*, attackers can learn privacy information about users and cached content by observing and measuring activities relevant to edge caches such as response time, power consumption, and return faults [2, 42, 71, 85]. Through the edge caching service, users can conveniently upload their content to ESs or download requested content from ESs. Due to the open accessibility of ECs [54], adversaries can easily access content cached by ESs. Adversaries can target a particular victim user by identifying content requested by the victim. The attacker may know the victim's content consumption habits or other specific characteristics to distinguish the victim from other users. One of the main types of cache side-channel attacks is *cache-timing attacks*, which allows attackers to determine whether specific content has been cached by comparing response times. Previous works (e.g., [2, 42, 71]) have explored

cache-timing attacks in edge caching systems. An attacker can conduct the precise timing measurement to distinguish cache hits from misses, which can identify what content is cached at the ES. A cache hit means that a nearby user has requested the content (or has a high caching value), whereas a cache miss means that the content has not been requested (or has been ejected from the cache). A knowledgeable attacker can further determine whether the request is served by the provider or by a router somewhere along the provider's path [42]. The main risk entities to launch cache side-channel attacks include malicious end devices [2, 42, 85] and external adversaries [71, 113].

**3.1.4 Cache Tampering Attacks.** A *cache tampering attack* is a form of cyber-attack in which an adversary aims to alter content stored at an EC to gain unauthorized access, introduce illicit content, and disrupt the caching system's regular operation. Within an edge network, a caching server offers a temporary storage area, holding frequently accessed content to expedite distribution. However, cache tampering attacks can transpire when an attacker modifies content cached in the ES or deceives the user to gain unauthorized content. The main risk entities to implement cache tampering attacks include ESs [61, 75], malicious end devices [17, 61], and external adversary [17, 61].

A typical instance of cache tampering attacks is *cache poisoning*, where an attacker manipulates a CDN or ES's cache to store and deliver malicious content or information [17, 75, 113]. For example, an attacker can exploit the vulnerability of the caching system by requesting a legitimate image with a specially crafted HTTP header. This header may contain malicious code that tricks the cache into storing a different image the attacker controls rather than the legitimate one. The next time a user requests the original image, it will instead receive the attacker's image, which could contain harmful content such as malware or phishing links.

A variant of the cache tampering attack is the *cache deception attack*, wherein an adversary gains access to private information by misleading and influencing a privileged user [24, 49, 50]. This process consists of two primary steps [24]. Initially, the attacker prompts the privileged user to request sensitive content and cache it in the ES. Subsequently, the adversary submits an identical request to the EC and retrieves the sensitive content. For example, in named data networking [38], an attacker creates a URL request targeting a victim user's private content by attaching a tag of a widely used image. The victim is then enticed to make that request using its privilege [49]. Upon retrieval, the cloud server disregards the invalid suffix and returns legitimate privacy content. The caching node retains the privacy content as the popular image's content. In this manner, the attacker can make the same request to access the identical privacy content in the EC, enabling them to acquire private content they are not authorized to access, potentially resulting in the victim's private content being leaked [49, 50]. The preceding kinds of cache tampering attacks give rise to unbearable privacy risks for users in edge caching systems.

## 3.2 Mitigation Methods to Preserve Privacy in Edge Caching Systems

In the following subsection, we will provide a concise introduction to a range of methods that can effectively mitigate privacy leakage in content caching systems, which can be mainly classified into four types of methods: noise-based methods, **trusted distributed computing (TDC)**, cryptology-based methods, and other approaches. The specific solutions corresponding to each privacy mitigation approach are detailed in Sections 4 through 6. For easy reference, we also present a classification matrix for the solutions introduced in this survey based on countermeasures and privacy data in the realm of edge caching in Table 2 in the appendix.

**3.2.1 Noise-Based Methods.** *Noise-based methods* represent the most prevalent approaches for preserving privacy within edge caching systems. These methods introduce disturbances to the real and genuine information before its exposure and interaction, effectively safeguarding privacy.

Within the domain of edge caching, three specific types of methods are commonly employed: **differential privacy (DP)**, confusion, and anonymization.

*DP* is a data-sharing technique that allows data owners to share only some statistical characteristics of a database while withholding individual-specific information [2, 71]. There are two ways to add noise in the DP mechanism. The traditional one is to add noise to the public database at the time of data release. However, the data collection agency is not always reliable, and thus a **local differential privacy (LDP)** mechanism is also leveraged by data owners to distort original data before submitting private data. The use of DP in edge caching systems can introduce distortion to the actual user or content information during the collection or release of sensitive data. DP is introduced to protect request traces [71, 81, 111, 116, 122] personal information [105, 124], and machine learning models [101] in edge caching systems.

*Confusion* mainly has two ways to enhance privacy in edge caching. The first one is cache obfuscation (e.g., proactive cache [55, 61], off-path cache [85], and request hit delay [42]), which can be used to protect users' requests when retrieving the content from monitoring or timing attacks in an untrusted or semi-trusted network environment. The second one is spatial confusion [3, 25, 110], which is to protect the location information when users enjoy LBSs. For instance, many pseudo requests for POIs can be attached to the genuine request when retrieving content from the EC.

*Anonymous* methods are the last category of privacy risk mitigation measures. Anonymity is the act of not being named or using an alias, as opposed to the act of having a real identity [18]. In particular, a set of public data satisfies  $K$ -anonymity if the information of any entity cannot be distinguished from at least  $K - 1$  other entities. The  $K$ -anonymity method is often used to enhance geographical [30, 56, 99] and personal privacy identity information [18, 67] in edge caching systems. In addition, the anonymity group technology is used in protecting users' identity information [53, 95, 96, 113].

**3.2.2 TDC-Based Methods.** TDC methods encompass three primary mitigation frameworks—*FL*, **secret sharing (SS)**, and blockchain technology—to safeguard privacy in the context of edge caching.

*FL* is a distributed machine learning technique that trains a learning-based algorithm across multiple decentralized devices or ESs locally holding data samples without exposure [9]. The *FL* framework is one of the most essential methods to preserve private data during the machine learning process. It is common that the *FL* framework [40, 45, 83, 84, 100, 102, 103] trains learning models by exposing model parameters or gradients. Instead, traditional machine learning methods need to collect raw data for the learning process. However, model parameters or gradients are also private assets of users since attackers can infer and recover users' private information from exposed model information. In addition, model information may have significant economic benefits, which will compromise the self-interest of model owners if they are exposed directly. A number of works [11, 16, 80, 101] have contributed to upgrading the *FL* framework by injecting noise or other interference to model information prior to exposure.

*SS*, also known as secret splitting, is a kind of secure multi-party computation and storage method in which each party gets a part of the secret, called a *secret share*. The secretly shared information cannot be recovered unless a sufficient number of secret shares can be collected. A single share cannot restore the original secret. For example, the  $(t, n)$ -threshold scheme is the most straightforward SS scheme. In this scheme, there are a total of  $n$  players, each receiving only one secret share. The secret can be recovered if at least  $t$  players cooperate, where  $t$  is the safety threshold parameter. In edge caching scenarios, secret data may include private content generated and stored by users and historical data required for edge caching decisions (e.g., request traces [2], user preferences [66], and content popularity information [6]). While introducing SS may increase

computational load, it significantly raises the cost for attackers attempting to obtain private information from the edge, reducing the risk of data breaches. Moreover, it enhances the fault tolerance of distributed caching systems.

*Blockchain* is a technical solution that does not rely on third parties to carry out network data storage, verification, transmission, and communication through its own distributed nodes. The blockchain mechanism can automate the following four steps. First, when a new blockchain transaction occurs, all participants can competitively record that transaction as a data block. Second, following the rule of consensus, most participants on the blockchain network must vote for a valid recorded transaction. Depending on the type of network, the consensus mechanism of agreement can vary but is typically established at the start of the network. Third, once participants have reached a consensus, transactions on the blockchain are written into blocks appended to a cryptographic hash that links blocks together as a chain. Fourth, the blockchain system finally updates and broadcasts a copy of the latest ledger to all participants. Blockchain can be used to enhance the protection of user preferences [61], personal information [19, 38, 44, 77], and machine learning data [16] in edge caching systems.

**3.2.3 Cryptology-Based Methods.** *Cryptology-based methods*, as a vital category of mitigation approaches, play a significant role in preserving privacy within edge caching systems. These methods employ cryptographic techniques to safeguard sensitive content or information, ensuring confidentiality, integrity, and authentication. Within the realm of edge caching, three specific types of methods are leveraged: encryption communication, **homomorphic encryption (HE)**, and **private information retrieval (PIR)**.

*Encryption communication* is divided into two steps to protect the security and privacy of communication data. The first step is to encrypt communication data as follows. The sender encrypts the content by an encryption algorithm and the receiver's public key to obtain the ciphertext. The receiver, once getting the ciphertext, conducts decryption through the decryption algorithm and the private key to recover the original data. Encryption communication is commonly used to protect the security of user request records and other data in Internet communications. There are three main approaches for encryption in edge privacy-enhanced caching systems. One is symmetric encryption, which mainly uses DES (Data Encryption Standard, AES (Advanced Encryption Standard) [59, 104], or SE (Searchable Encryption) [17]. Second, asymmetric encryption mainly includes RSA (Rivest-Shamir-Adleman) [94], ABE (Attribute-Based Encryption) [59], and ECC (Elliptic Curve Cryptography) [14, 113]. Finally, there are hashing algorithms [94, 95, 96], which are sometimes used in blockchain [16, 38]. However, there are also three significant concerns with the use of cryptographic methods in edge caching systems. First, due to the existence of encryption, third-party ECs often cannot directly use encrypted requests to retrieve related content, which may lead to the unavailability of ECs. Second, introducing encryption technology may pose computational pressure on the resource-constrained edge and end devices. Last, encryption communication may fail to prevent record privacy from CPs or service providers, who have the key to decrypt request information. Therefore, how to introduce cryptology-based techniques into edge caching systems is still a challenging problem. In addition, as a special communication encryption method, the digital signature [11, 32, 34] is often used in edge caching systems to verify user identity and data reliability.

**HE** is a form of encryption by which each party co-computes the result of a specific objective function concerning their private data without a **trusted third party (TTP)**. Each party cannot unveil private data from other parties even if the computation is completed. In other words, it allows a participant to perform operations such as searching and multiplying encrypted data to

produce correct results without decrypting it during calculation. HE can be used to protect user preferences [14] and information [34] when searching the EC.

PIR is mainly used to protect a user's request record information [36, 75] in the edge caching system. When obtaining sensitive data, request records likely expose important privacy information of users. PIR can help users with query needs to complete private data retrieval from the EC under the condition that the query privacy information is not leaked. In other words, the PIR technology can prevent attackers from obtaining precise query information and content items in cache retrieval or other sensitive queries. At the same time, PIR can let users obtain desired private content.

**3.2.4 Other Methods.** *Optimization-based methods* and *access control* are two of the main approaches to enhancing the effectiveness of privacy protection in edge caching systems. In optimization-based methods, metrics such as privacy exposure [4, 71] and credibility [10, 94, 121] are mathematically modeled. The quantified metrics are then regarded as the objective function or constraint variables of the cache optimization problem. Finally, the optimal privacy protection decisions are deduced by solving the optimization problem [27, 28, 69, 93]. Access control is an enforcing control method that allows or denies a user's access to a specific network resource (e.g., private content in the EC) based on the user's account or group. Without a defined authorization mechanism, access to system resources will have no restrictions, and thus illegal device operations can be easily launched. The EC can implement strict access control to filter out unauthorized or illegal accesses into the caching space for privacy protection. Access control methods have been applied to protect personal information [17, 38, 113] and content privacy [95, 96] in edge caching systems. In the next section, we dive into the details of defense methods for protecting each type of sensitive information.

**3.2.5 Summary.** In conclusion, privacy-preserving methods in edge caching systems can be broadly classified into four categories. Noise-based methods are among the most prevalent techniques for safeguarding privacy. These methods are particularly effective in preventing monitoring attacks and cache side-channel attacks but may negatively impact the utility and performance of edge caching systems. For instance, DP is employed to add noises to data before or during its release, protecting user request traces, personal information, and machine learning models from privacy breaches. However, due to the limited privacy budget, strategic account [1] or allocation mechanisms [88] are required to mitigate the adverse influence of noises on utility. Noise-based methods typically introduce an acceptable level of computational overhead, providing adaptive protection in real-time edge systems where capacity is limited.

TDC-based methods refer to techniques that organize distributed devices for collaborative computation and storage while ensuring data privacy and security, such as FL, SS, and blockchain technology. FL is valuable for protecting the privacy of machine learning models by enabling decentralized training without exposing raw data. These methods help mitigate the risk of data mining attacks, where adversaries might otherwise exploit original request traces or model parameters to infer private information. However, TDC-based methods, regarded as a form of adaptive protection, also need integrate with some strict privacy-preserving methods and may introduce communication and computational complexity due to the need for synchronization and consensus across multiple nodes, which can affect scalability and deployability.

Cryptology-based methods utilize cryptographic techniques to ensure the confidentiality and integrity of sensitive information. For example, HE allows computation on encrypted data without needing to decrypt it first, protecting user preferences during data processing. In addition to preventing request information from being monitored by external attackers, these methods are crucial

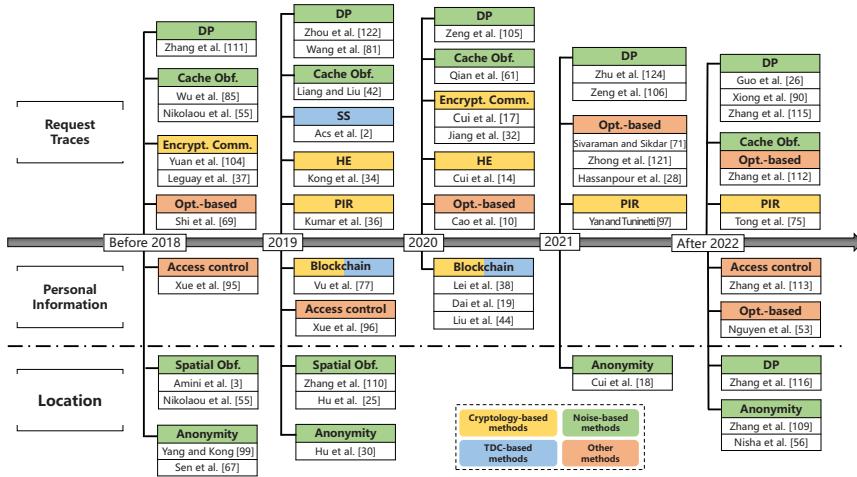


Fig. 4. A brief timeline of solutions aimed at enhancing user privacy, including request traces, personal information, and location, in the EC. Each solution is accompanied by its main mitigation approach.

for preventing cache tampering attacks, where attackers might alter cached content to gain unauthorized access or disrupt system operations. Despite offering stringent protection, cryptology-based methods generally incur high computational complexity, particularly in scenarios involving encryption and HE, which can become bottlenecks for resource-constrained edge devices.

Last, other methods, including optimization-based techniques and access control, enhance privacy by mathematical modeling and enforcing access restrictions to sensitive resources. The complexity of privacy-preserving methods in edge caching systems varies significantly depending on the approach. Optimization-based methods involve solving complex mathematical models, with computational intensity heavily dependent on the formulations and solving algorithms. This complexity may be particularly high when optimizing multiple privacy metrics simultaneously or under complex constraints. Therefore, selecting appropriate privacy-preserving methods requires a careful balance between the desired level of privacy and the performance cost. For easy reference, we also present a classification matrix for the solutions introduced in this survey based on countermeasures and privacy data in Table 2 in the appendix.

## 4 Enhancing User Privacy in Edge Caching Systems

User privacy is the most important privacy in edge caching systems, which has attracted tremendous research efforts dominating the research on privacy preservation in edge caching systems. We discuss these defense methods based on three types of user privacy: request traces, personal information, and location.

### 4.1 Privacy of Request Traces in the EC

*Request traces* are the most critical privacy information in the EC, from which adversaries can obtain user preferences [81]. We summarize methods to protect user request records from four aspects, which are noise-based methods, TDC-based methods, cryptology-based methods, and other methods. A brief timeline of solutions for enhancing the privacy of request traces is presented in Figure 4. The solutions for enhancing other user privacy, such as personal information and location, are also summarized in Figure 4 for the sake of brevity.

**4.1.1 Noise-Based Methods.** The initial class of methods to protect request records are *noise-based methods*, which can be categorized into two main approaches. The first approach involves adding noises generated by mechanisms such as DP to protect information [26, 81, 82, 105, 106, 111, 122, 124]. The second approach includes cache obfuscation methods (e.g., proactive cache [55, 61], off-path cache [85], and request hit delay [42]) to protect users' requests from monitoring or timing attacks in untrusted or semi-trusted network environments. We will elaborate on these methods in the following sections.

**Differential Privacy.** CPs often utilize edge caching nodes at the edge network and collect users' private access records to predict user preference to improve delivery efficiency. However, directly collecting users' profiles can lead to privacy breaches. Additionally, in highly dynamic scenarios, the entities of the EC (e.g., ENs [122] and ESs [124]) collect user request records in real time and make dynamic decisions to improve the efficiency of edge caching. This real-time data collection process also poses a risk of privacy leakage, where DP-based methods can be employed to mitigate the risk.

Zhou et al. [122] proposed a privacy-preserving and online distributed multimedia content retrieval system. Each EN in the system is modeled as an online learner to exploit user requests with a context that includes their background information (e.g., age, gender, location, social profile, and query criteria). The ENs can collaboratively make multimedia content recommendations and cache at the edge network. When an EN needs extra context information to make a retrieval scheme, the TTP sends noisy records to ENs by deploying the DP mechanism. A trust mechanism is also proposed to identify and remove malicious ENs. Zhu et al. [124] studied the tradeoff between privacy protection and caching efficiency in edge caching systems. When a user generates a content rating vector, Gaussian noises are added to the original rating vector, then the distorted rating vector is transmitted to the ES for privacy protection. In the global aggregation information stage, the ES calculates the eigenvalues and eigenvectors of collected data based on the lightweight-level calculation algorithm. Then, the ES broadcasts the results to all users. Xiong et al. [90] presented a novel network traffic shaping framework for protecting privacy in IoT networks by integrating DP with constrained optimization. They developed a tunable DP model that shapes encrypted IoT traffic to protect against monitoring attacks, particularly eavesdropping on packet sizes and timing. This approach not only safeguards IoT traces from privacy breaches but also enhances the resilience of IoT systems against traffic analysis attacks by dynamically adapting to changing network conditions and heterogeneous user demands.

In collaborative edge caching, managers exchange sensitive information, such as user records or preferences [105, 122], and routing records [105], to improve caching efficiency. However, protecting privacy often in collaborative ECs may rely on a centralized TTP, which is challenging to obtain in practice and places more pressure on network bandwidth. Moreover, if the centralized TTP is attacked, it may pose a more serious privacy breach risk. Zeng et al. [105] proposed a distributed method to develop network caching and routing strategies for SBSs. The scheme adds a DP noise in the routing information (i.e., the portion of the requested content served by each SBS) during the exchange process to protect the privacy of SBSs and **mobile users (MUs)**. It defines an optimization problem that minimizes the global cost, which is solved by a distributed protocol. Guo et al. [26] introduced a blockchain and DP-based decentralized edge-thing system for privacy preservation and fair utilization of edge computing resources. The proposed system employed blockchain technique to deal with transactions and smart contracts' tempering issues caused by the malicious auctioneer node. Moreover, an exponential mechanism-based DP is applied to the double auction scheme to tackle the inference attack on auction results saved in the blockchain.

Hits on the user's local cache can provide the best service experience for users. However, it is challenging for end devices that rely on a user's personal historical information to make accurate prefetching decisions solely. Collaborative efforts between users are necessary, but such information exchange is risky, and the recorded history must be protected when disclosed. Wang et al. [81] presented a mobile video prefetching strategy based on DP and distributed online learning algorithms. They formulated the prefetching problem as an online optimization problem considering user preferences, video popularity, and social connections. The problem is then decomposed into two subproblems, which are solved and swapped at each terminal by a distributed method to obtain the optimal global solution. The DP mechanism is added in exchanging user-sensitive information during each round of iteration to protect user privacy.

**Cache Obfuscation.** In an ICN, users can directly access desired content from edge routing nodes. However, edge routing nodes are often vulnerable to cache side-channel attacks, which can result in the exposure of requested record privacy. Liang and Liu [42] designed a method to defend against timing attacks in Content-Centric Networks. According to the privacy protection degree for requested content and the honesty degree of requested nodes, evaluated by the historical information, the caching node calculates the delay in responding to requests to defend against timing attacks. Further, Wu et al. [85] designed a multi-path caching strategy for ICN based on random linear network coding. The strategy encodes different video chunks into the same block for efficient content delivery. When the block is delivered along the path, it can only serve all routing nodes with related video chunk requests and keep unavailable to irrelevant nodes. It adopts a random forwarding method that increases the diversity of routing paths, thereby increasing the size of anonymity sets and the cost of inferring user privacy.

In addition, proactive caching of redundant and obfuscated content at the edge can interfere with an attacker's ability to access the user's actual request records. Qian et al. [61] proposed a privacy-aware content caching architecture for cognitive IoV networks with proactive caching and blockchain technology. In this system, RSUs and smart vehicles can cache content in advance, which can provide the cached content in the form of a broadcast to meet the content needs of other vehicles. Therefore, a vehicle only needs to obtain content from broadcast data without further requests, which can reduce user privacy exposure. At the same time, blockchain technology is introduced to ensure a more secure and reliable transaction mode to guarantee the reliability of the content. Additionally, Nikolaou et al. [55] proposed two cache placement strategies for the joint caching of users. The first strategy considers the graph network structure between user terminals, and the second one focuses on the workload change of the server. However, transmitting requested videos between clients will leak privacy for both sides. The requested user proactively fetches and caches obfuscated content. At the same time, the server adds randomly obfuscated addresses when sending feasible retrieval address lists to reduce the risk of privacy exposure.

**4.1.2 TDC-Based Methods.** The second category of *TDC-based methods* aim at safeguarding request records primarily comprises SS, a secure multi-party computation technique, that can effectively prevent attackers from acquiring valued request records. Acs et al. [2] proposed two timing attack defense methods for the edge router cache in the ICN. For interactive traffic-type communication, random naming and SS are used for privacy protection to prevent attackers from obtaining specific traffic information. In view of the content distribution traffic, a method of increasing artificial delay is proposed to protect privacy, and a certain delay is added to the private content that is hit by the router cache to prevent adversaries from determining the hit status of private-sensitive content.

**4.1.3 Cryptology-Based Methods.** *Cryptology-based methods* have been widely used to protect the security and privacy of user request records and other information in Internet communications. However, there are also three challenging problems when using cryptographic methods to protect the privacy of request records in edge caching systems. First, due to the existence of encryption, third-party ECs probably cannot directly use encrypted requests to retrieve related content, leading to the unavailability of ECs [37, 104]. Second, introducing encryption technology may pose heavy computational pressure on the resource-constrained edge and end devices. Last, cryptology-based methods fail to prevent the leakage of record privacy from CPs or service providers, who have the key to decrypt requests. Therefore, how to apply cryptology-based techniques to edge caching systems is still a challenging problem.

**Encryption Communication.** To prevent the monitoring of users' request records by ISPs, efforts have been made to encrypt request records and the corresponding transmitted data using encryption algorithms while ensuring the availability of the cache within the ISP. Yuan et al. [104] designed a system to achieve efficient encrypted video delivery in the ISP network. The content cached in the network is encrypted and distributed in the ISP network. This system can efficiently and safely locate and retrieve related content from the ISP network with a proposed encrypted content fingerprint index for a given encrypted request.

To improve privacy in the CDN, Cui et al. [17] proposed a novel encrypted method that combines SE and a multi-CDN strategy to achieve both content delivery performance and security in edge CDN nodes. The work introduces the SE method to realize content security and searchability. In addition, a semantically secure algorithm is used to encrypt user requests so that the same query can correspond to different request content. To further protect user preference privacy, a one-time nonce will also be used for secondary encryption, which will be transmitted together with the content transferred between CDN node clusters. For each request, the node must receive the nonce to search, and after the search hits, the nonce must be regenerated and re-encrypted before continuing to deliver the content.

**Homomorphic Encryption.** In previous works, HE has been introduced to protect the privacy of vehicles' request records in the IoV while collaborating with RSUs to improve the efficiency of edge caching. Cui et al. [14] proposed a cooperative download scheme in the IoV network, considering the security and privacy protection of request traces. This scheme uses edge computing architecture to reduce transmission delay. It uses lightweight encryption methods, such as ECC, the Tesla broadcast authentication, and additive HE, to protect user privacy and content security. The strategy proposed in this work is composed of two phases: the non-accelerated phase and the accelerated phase, the details of which can be found in the work of Cui et al. [14].

Kong et al. [34] utilized an invertible matrix to construct multiple content requests sent by different vehicles such that the RSUs can recover each request without being associated with a specific car. Specifically, when a vehicle needs to initiate a request, it will first generate a  $k * k$  random invertible matrix and send secret information required for HE to  $k$  vehicle users within a unified range. Then, in the response, a collaborative request group is randomly selected for the requested vehicle. Other vehicles in the group first generate the requested information according to the Paillier HE algorithm and send it to the RSU, returning the HE information to the requested vehicle. That vehicle completes the corresponding HE according to the returned information and the invertible matrix. Finally, it sends the encrypted request to the RSU to retrieve the private content without exposing its privacy.

**Private Information Retrieval.** By utilizing PIR methods, users are able to obtain the content they desire while preventing potential leaks of their private interests. Kumar et al. [36] were the first to introduce a PIR strategy based on encoding cache into wireless edge caching.

Erasure-correcting codes are used to encode cached content, and different bit rates can be selected for videos with varying popularity to conserve backhaul bandwidth usage. Additionally, the scheme is based on general Reed-Solomon coding to safeguard user privacy from SBSs that may collude with one another. Furthermore, ensuring the integrity of content in the EC is essential for maintaining a stable edge caching system. This is particularly important because edge devices owned by individuals or small organizations are susceptible to cache tampering attacks and internal hardware failures. However, verifying the integrity of the content can compromise its privacy, especially when third-party verifiers are involved. To address this issue, Tong et al. [75] proposed an integrity-checking protocol for edge storage based on provable data possession to verify the integrity of cached content on a single EN. The protocol employs a PIR scheme and homomorphic verifiable tags to prevent the disclosure of sensitive information (e.g., user request traces, edge download schemes, and private content) to verifiers.

**4.1.4 Other Methods.** Other methods, such as *optimization-based methods*, are also introduced to enhance the privacy of request traces or user preferences in EC. Sivaraman and Sikdar [71] used game theory to formulate an off-path and cooperative caching problem in the edge of the ICN, where users can choose their optimal routers at the edge network to cache content. Constraints in the problem include network latency, caching cost, and the amount of exposed user privacy. Two different privacy measures (i.e., mutual information and DP) are used as constraints in the work. Finally, it is proved that a Nash equilibrium point exists in the game, which can be solved by an iterative method. In addition, to mislead adversaries eavesdropping on ECs, Hassanpour et al. [27, 28] proposed caching approaches aimed at enhancing privacy and reducing communication costs in edge networks. The solution presented one work by Hassanpour et al. [27] employs an  $\epsilon$ -constraint optimization approach to balance the tradeoff between minimizing the average delivery load and maximizing context-oriented privacy. By optimizing cache placement probabilities, the approach in another work by Hassanpour et al. [28] utilizes chunk-based **joint probabilistic caching (JPC)** to increase adversarial errors while maintaining the desired privacy levels. Furthermore, to address the exponential growth of the feasible solution set in the JPC optimization problem, they proposed a scalable JPC strategy to solve the linear programming optimization problem efficiently.

Furthermore, Cao et al. [10] studied the reliable and efficient performance of multimedia transmission services between base stations and MUs through a two-stage joint optimization. In the first stage of optimization, a service reliability evaluation mechanism is designed to evaluate the credibility of the base station to ensure the security of user privacy information. Then, the price and reliability competition among base stations and the strategic interaction of all players are modeled by the Stackelberg game [29]. A resource allocation problem is further proposed in the second stage to coordinate multiple MUs serving on the same base station. The potential game model is used to improve the transmission service performance. Additionally, Shi et al. [69] proposed a model for the cache placement problem in wireless edge caching, considering a multi-attacker scenario where both benign users' and attackers' locations follow a homogeneous Poisson point process. An optimization problem is formulated to determine the probability of each caching file, considering the average probability of successful eavesdropper attacks and transmissions at the wireless edge network. Finally, the genetic algorithm is used to maximize the secure transmission performance of the system.

## 4.2 Privacy of Personal Information in the EC

*Personal identity information* is also sensitive in the network, which can be used by the EC for carrying out sensitive operations such as permission control and cache admission control.

However, excessive disclosure of users' personal identity information makes it convenient for malicious nodes and attackers to spam users with advertisements and recommendations and attack ESs by polluting cached content.

**4.2.1 Blockchain-Based Methods.** Previous works mainly employ blockchain to protect users' identity information [19, 38, 44, 77]. Specifically, Vu et al. [77] proposed a **blockchain-based CDN (B-CDN)** architecture for content delivery, which enables anonymous operations on users. The B-CDN leverages intelligent contracts to maintain the blockchain and provide CPs with users' registration and subscription functions while ensuring user privacy. Additionally, the B-CDN can reduce the cost of CP management by utilizing a public database of requested traces, which allows CPs to estimate users' preferences with virtual identities and maximize the efficiency of their caching services.

The **Named Data Network (NDN)** is a variant of the ICN, where content can be retrieved by the content name. Lei et al. [38] introduced a blockchain-based security architecture for improving the security and privacy of NDN-based vehicular edge computing systems. This work deploys blockchain nodes in ESs and ISP nodes, where a delegated consensus algorithm is designed to enhance the efficiency of the blockchain. A three-layer management framework and an access control strategy are proposed for key management based on blockchain verification and vehicle attributes, respectively. A resource requester needs to prove to blockchain consensus nodes that it satisfies the access condition according to the access policy of the resource owner. Dai et al. [19] designed a content caching mechanism based on the permissioned blockchain technology to address the problem of privacy and security in the vehicle edge computing network. A new block validator selection method is proposed to achieve a fast and efficient blockchain consensus mechanism. In addition, this work presents a deep reinforcement learning based vehicle content caching algorithm. Liu et al. [44] designed a decentralized caching framework empowered with blockchain credentials to tackle the challenges of content data verification and edge device authentication. In the designed system, it is possible to trace each transaction at an active edge network without a central manager. A cache order matching technique is devised to use the cache resources efficiently. Further, data integrity verification is done with the help of a content trading mechanism that helps data sharing among the edge devices of the edge network and ensures the efficiency of trading in the edge caching system.

**4.2.2 Other Methods.** The *access control* is also exploited to protect users' identity information [53, 95, 96, 113]. Xue et al. [95, 96] proposed SEAF, a secure and efficient network access framework for cache resources at the edge of ICN. SEAF provides several security and privacy features, including content confidentiality, user privacy protection, user privilege revocation, countability, and efficiency. In SEAF, routers at the edge network authenticate user requests to separate access control from content provisioning. Only authenticated requests can enter the network; thus, authorized users can only access the bandwidth and cache resources inside the ICN. Meanwhile, to protect privacy, users can verify their identity to the edge router by generating a valid group signature, thereby maintaining users' anonymity to the edge router. Zhang et al. [113] focused on the security issues of cache-based software-defined networks, using the Tesla protocol to achieve fast authentication of the cache of vehicles and fog nodes. Besides, the Pedersen commitment mechanism is used to directly authenticate vehicles and fog nodes without exposing user identity privacy. Considering the limited computing power and delay-sensitive characteristics of the IoV, the author designed a set of cryptographic mechanisms supporting batch verification.

### 4.3 Privacy of Location in the EC

The *location* information is a kind of critical privacy of a high value, including moving trajectory [39, 86, 116], spatial coordinates [110], and other unique features [116]. Noise-based methods comprise the primary class of techniques employed to enhance location privacy as illustrated in Figure 4.

**4.3.1 Noise-Based Methods.** As such, noise-based methods are mainly introduced to protect location privacy, including geographic DP [116], spatial confusion [3, 25, 110], and  $k$ -anonymity [18, 56, 67, 99, 110], among others.

**Differential Privacy.** With the increasing mobility of users and the constant threat of malicious attacks from third parties, there is a growing risk of privacy breaches in mobile edge caching. To address this issue, Zhang et al. [116] proposed a DP-based method for improving the video QoE for MUs while protecting their location and preference privacy in mobile edge caching. The proposed scheme utilizes a privacy-preserving approach for computing the location transfer model and aggregating user preferences, achieving a balance between caching service efficiency and privacy protection at mobile edge networks. Specifically, the Laplacian perturbation model is employed to protect users' location and preferences when submitting their information. Based on the perturbed information, mobile edge caching nodes can evaluate the popularity of videos in the user's area, and Q-learning [73] is employed to achieve cache optimization goals combined with transcoding technologies.

**Spatial Obfuscation.** Amini et al. [3] were one of the first to utilize devices' cache to protect users' location information, where location-based content can be periodically prefetched to devices in large geographic blocks before they are actually consumed. When content has been cached in a user's local area, the user can access it directly on their device without needing external network services. This can effectively reduce privacy exposure risks for the user.

Additionally, privacy protection can be achieved through a distributed collaborative cache that forms anonymous user groups within the vicinity. Zhang et al. [110] proposed a multi-level caching strategy to reduce the number of users directly requesting LBS from the **local service provider (LSP)**. In turn, users can obtain the required services from the local cache, surrounding neighbor caches, and trusted anonymizers. In this way, the interaction with untrusted LBS is reduced and privacy exposure is mitigated. When the request is lost, it has to request the LSP by generating a stealth zone and making a request to the LSP. The anonymizer will select the optimal  $K$ -space anonymity to request content according to the prediction result (considering a user's future geographic location, the caching contribution rate of each unit, and the freshness of the content in the unit). However, the high communication overhead and computational energy consumption of users collaborating as a group pose problems in protecting privacy. Moreover, the introduction of centralized anonymizers is vulnerable to attacks, and if breached, all users' private information may be compromised. To address these limitations, Gu et al. [25] proposed a method that employs the trusted ESs to preprocess user requests and blur their location information during the snapshot query (i.e., one-shot query) of their POI. The ESs cache the requested POI for further query, thus minimizing the number of queries exposed to LBS providers and potential attackers. Additionally, in continuous queries, fuzzy prediction queries are generated and correlated with the actual query to enhance the queries' utility while interfering with attackers.

**Anonymity.** The utilization of cache in edge devices, such as user devices [18, 56, 99], ESs [67], and RSUs [30], can keep users' transparency from LBS providers by reusing the users' POI within a specific region. This approach allows users to access the cached POI directly at the edge network instead of relying on remote LBS service providers. Additional privacy protections (e.g.,  $k$ -

anonymity [30, 99, 109],  $l$ -diversity [18], anonymity groups [56, 67]) are exploited when resources have to be obtained from LBS providers. As a result, the likelihood of exposing sensitive location information to the service provider is reduced.

Zhang et al. [109] devised a CDKA (Caching-based Dual  $k$ -Anonymous) mechanism to preserve location privacy. CDKA uses double anonymity and multi-level caching to reduce communication overhead while providing location privacy. For this, an ES is used to intervene between the user and the LBS server, and location privacy is ensured by making mobile clients and ESs anonymous. The proposed mechanism is assessed for computational efficiency, communication overhead, and cache hit ratio. Additionally, dealing with vehicles' high-speed movement characteristics in vehicular networks, Hu et al. [30] designed a privacy protection algorithm combining proactive caching and the  $k$ -anonymity method. When a vehicle user requests a specific POI, it needs to send  $k - 1$  obfuscated requests simultaneously. Besides, the corresponding request content will be obtained through multiple passing RSUs to protect the user's location information, including factual geographic and POI.

Moreover, at the edge of the wireless network, Sen et al. [67] proposed a double cache strategy to deploy a pair of caches for each region. Cache  $A$  records previous request results of users in the region, and cache  $B$  caches all user requests and maintains cooperation between users. When querying private content, a user queries cache  $A$  first. If it is not hit, the request will be redirected to cache  $B$  for conversion. Finally, the converted request is sent to other users within the region to request LBS together, and the received results may be maintained in cache  $A$  for further queries.

To further prevent users' location and personal information from being accessed by untrustworthy EC and malicious users, Nisha et al. [56] proposed a caching scheme called the *Group Collaboration Scheme* to request POI combining with spatial obfuscation. In this scheme, users who need to find a POI in a specific area will modify the requested area according to the proposed random area obfuscation algorithm and then register with the group authenticator to obtain virtual group identity information and cooperative anonymous user groups. The collaboration is one-time, and the anonymous group changes as the user moves. Users with request requirements will cooperate with nearby users to query whether the cache of other users in the anonymous group meets the request requirements. If the request POI is unavailable in the user group, the required content will be requested in the name of the anonymous group.

**4.3.2 Trusted Distributed Computing.** To enhance the QoS, CPs collaborate with ISPs to deploy edge caching resources as close to the users as possible. ISPs can support the EC by placing virtual servers at the network's edge and assigning them to CPs. However, CPs only possess the request records of users, whereas ISPs only have access to their geographic location information. In the caching process, CPs do not want to disclose all the requested information to the ISPs, and vice versa. To deal with this challenge, Andreoletti et al. [5] proposed a secure multi-party computation protocol to facilitate cooperation between ISPs and CPs without requiring either party to disclose sensitive information. The protocol enables ISPs to obtain the number of requests for specific video content in a given area at a low computational cost. Once the ISP has this information, it can deploy virtual servers efficiently, and the CP can use these virtual servers to place the EC, thereby minimizing the number of hops for content delivery and reducing communication delays.

Despite the comprehensive introduction of major solutions, our discussion is not exhaustive. Thus, we provide a supplementary introduction in Table 3 in the appendix, briefly introducing other solutions to protect user privacy in edge caching systems that have not been discussed in detail in Section 4.

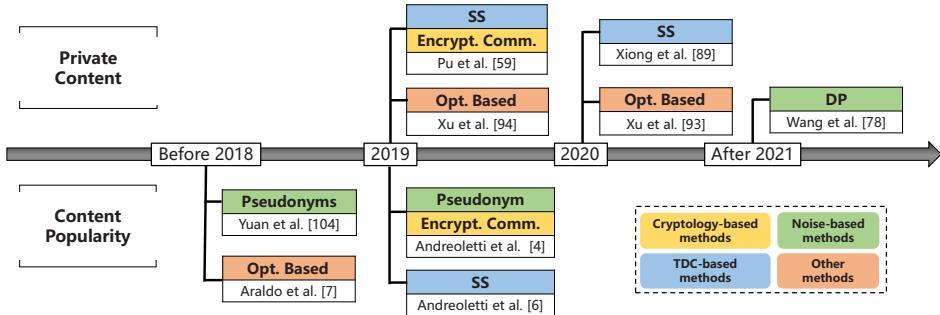


Fig. 5. A brief timeline of solutions for enhancing content privacy, including private content data and content popularity.

## 5 Enhancing Content Privacy in Edge Caching Systems

In this section, we move on to discuss defense methods that can preserve the second type of sensitive information (i.e., content privacy) in edge caching systems. We present a timeline, as depicted in Figure 5, summarizing the methods employed to safeguard content privacy, encompassing private content data and content popularity.

### 5.1 Privacy of Content Data in the EC

Other than caching content for CPs, ESs are also able to cache private content generated by users. However, due to the presence of incompletely trusted ESs [59, 94] or malicious and unauthorized users [93, 94] at the edge network, stored content in the EC may face privacy leakage risks.

*DP-based methods* are used to upload local data in the network cache while preserving its privacy. For example, Wang et al. [78] proposed DP-DLCF (Differential Privacy-Preserving Peep Learning Caching Framework) to deal with the privacy leakage problem of private content in edge caching networks. The privacy budget is utilized adaptively to strike a tradeoff between the privacy and accuracy of the prediction. In the proposed technique, users upload their data after perturbing it with a randomized response technique based on LDP to preserve the privacy of their local data. Next, the neighboring base station accumulates the uploaded data and transfers it to the deep model for training. Moreover, the prediction accuracy of the model training is improved by the bootstrap aggregation algorithm.

*Cryptology-based methods* can also be leveraged in protecting the private content in the EC. Pu et al. [59] proposed a secure and privacy-aware content-sharing strategy to protect sharing data stored and delivered by incompletely trusted ESs. To ensure the secure sharing of content, the content generator first encrypts the content using the CP-ABE (Ciphertext-Policy Attribute-Based Encryption) algorithm and calculates its signature based on its private key. Additionally, by utilizing the public key cached at the nearest ES, the generator performs secondary encryption of the content to the nearest ES. When the ES receives the encrypted content from the content generator, it will first decrypt the content with its private key and check the security of the content. According to the SS scheme, the ES randomly divides the content into  $n$  parts and distributes the content parts to other  $n - 1$  ESs to store the content. The proposed scheme can effectively ensure the integrity and recovery ability of the content in case any EC node becomes offline. The SS method was also integrated by Xiong et al. [89] to design an edge-assisted privacy-preserving data-sharing framework for autonomous vehicles. This approach encrypts raw data into two ciphertexts, which are processed by two ESs. Additionally, a privacy-preserving convolutional neural network (P-CNN) was developed to ensure that the classification results are

identical to those of the original convolutional neural network model, without any data leakage. The framework effectively addresses threats such as potential data leakage and unauthorized access to private content by malicious vehicles or ESs.

*Optimization-based methods* are introduced to enhance the privacy of content caching in ESs. To prevent private content from leaking to the unreliable edges and make optimal caching decisions for MUs, Xu et al. [94] used the multi-leader and multi-follower Stackelberg game to model a multi-link cache scenario at the mobile edge network. In the scenario, **edge computing small base stations (ECSBS)** act as leaders and, first, set pricing strategies in a non-cooperative game. Then, a trust mechanism is proposed to evaluate the reliability of each ECSBS, which consists of two parts: direct trust degree and indirect trust degree. Based on the caching reliability and pricing offered by ECSBS, MUs can make their optimal caching decisions as followers. Additionally, Xu et al. [93] proposed a Stackelberg game model to encourage **edge cache devices (ECDs)** to provide secure caching services in both static and dynamic scenarios. The model takes into account the selfish and open nature of ECDs and employs a zero-payment mechanism to penalize ECDs that provide poor services. The optimal strategies for the CP and ECDs in a static game are analyzed, proving the existence of a unique equilibrium in the Stackelberg game. Besides, in dynamic games with incomplete information, the Q-learning algorithm is used to solve the problem.

## 5.2 Privacy of Content Popularity in the EC

Content popularity, which can be used as the key knowledge to improve caching efficiency, is business-critical information for the CPs and edge caching managers (e.g., ISPs). Due to the limited number of records in the service scope of the EC (e.g., serving a specific geographical location range or a particular network level), edge caching suppliers may require CPs and other edge caching entities to provide the critical content popularity information so that they can judiciously make caching decisions so as to shrink bandwidth consumption of the core network.

Andreoletti et al. [4] improved the solution proposed in the work of Yuan et al. [104] by allowing CPs to encrypt content and associate them with pseudonyms to prevent privacy leakage to edge caching managers. ISPs only count the occurrences of these pseudonyms to infer content popularity without examining the original content. The authors introduced the mathematical definition of privacy and studied the tradeoff relationship among privacy and hit rate, retrieval latency, and traffic load metrics. Additionally, Andreoletti et al. [6] proposed a protocol for spatial partitioning of ISP caches based on the popularity of different CPs' content, which aims to improve the QoS of edge caching services while protecting CPs' privacy of popularity information. The protocol employs the Shamir SS scheme for CPs to share the popularity information between the ISP and the regulator authority, which guarantees a fair subdivision of the cache storage and the preservation of privacy. The ISP can calculate the caching space requirement for each CP using the secret information, thus protecting CPs' privacy.

Similarly, Araldo et al. [7] proposed a caching space partitioning method that protects the popularity information of CPs while ensuring the efficiency of edge caching. The method divides the ISP's caching space into multiple slices and assigns each slice to different CPs using the stochastic dynamic cache partitioning algorithm. The algorithm takes an initial slice allocation as input and iteratively optimizes the slice allocation scheme by testing the cache miss rate of the allocation scheme in each round. However, unlike the partitioning method proposed by Andreoletti et al. [6], this method does not depend on the private information of CPs' popularity. Additionally, this architecture also supports a transparent cache of encrypted content deployed at the edge of the ISP network.

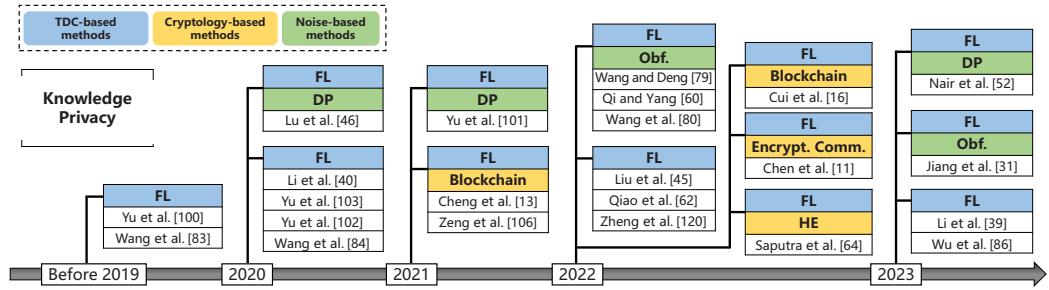


Fig. 6. A brief timeline of solutions for enhancing knowledge privacy.

## 6 Enhancing Knowledge Privacy in Edge Caching Systems

In this section, we discuss defense methods that can preserve privacy for the last type of privacy (i.e., knowledge privacy) in edge caching systems. All edge caching service providers have the motivation to extract knowledge for improving caching performance, which gives rise to the tradeoff between caching performance and privacy protection. Due to the high dynamics and complicated access patterns driven by users' interest [51, 98], it is essential to come up with intelligent edge caching algorithms to improve the caching performance. Machine learning based methods provide a feasible framework to extract user access patterns by exploiting collected datasets related to users, which may contain sensitive information. For example, video request access patterns are driven by users' interest in different locations [20, 47]. Users may keep dynamic moving [19], and their interests evolve over time [114]. Thus, it is necessary to make edge caching decisions based on features which can be extracted from localized and private user information by machine learning methods.

FL as a distributed machine learning framework is the most popular method to preserve knowledge privacy. FL trains a learning-based algorithm across multiple decentralized devices or ESs holding local data samples without exposing them. Additionally, we provide a comprehensive summary of the FL-based methods employed to safeguard knowledge privacy, presented in a timeline illustrated in Figure 6. As well, Table 4 in the appendix offers a detailed classification of these solutions based on the combination of methods used.

### 6.1 Enhacing Knowledge Privacy with FL Frameworks

The most common approach is to use an FL framework to train prediction models. Unlike traditional machine learning methods, FL does not collect raw data for model training [100, 102, 103]. This framework encourages models to be trained on local data, and all training works upload model parameters or gradients rather than sensitive raw data. Yu et al. [100] were probably the first to propose a learning-based proactive content caching method following the FL framework. This work proposes a hybrid filtering method based on the autoencoder to calculate the user-content similarity and predict the content of a user's interest. Yu et al. [103] also designed an FL-based proactive caching method for vehicular networks. Considering the high mobility of vehicles and dynamic content popularity in vehicular networks, RSUs integrate the mobility-aware cache replacement policy to make proactive caching decisions. Following the FL framework, the preceding three works [100, 102, 103] enable users to train machine learning models (e.g., autoencoder model) with their private datasets, locally and distributively, and upload trained models to the corresponding parameter server for aggregation.

Reinforcement learning can be realized in the FL framework to solve the complex dynamic control problem and mitigate the privacy leakage problem in edge caching systems [1, 40, 45, 62, 83, 84, 87] to improve the caching performance and privacy protection simultaneously. Wang et al. [83] proposed an “In-EDGE AI” system with deep reinforcement learning in FL. It delegates the reinforcement learning training task to the device side to protect the private dataset and brings more intelligence to edge systems. Liu et al. [45] proposed a privacy-preserving distributed deep deterministic policy gradient scheme to make caching decisions for the EC. At the same time, to preserve user privacy, the model only predicts content popularity by avoiding mining sensitive historical information. The model training process is completed by FL to prevent users from leaking privacy to ESs. Qiao et al. [62] proposed an FL-based proactive content caching scheme to shorten content retrieval latency and protect users’ private datasets. First, the edge computing architecture reduces energy consumption and transmission overhead. The problems of client selection and local iteration round selection in the FL process are modeled as an MDP, which is solved by the deep reinforcement learning algorithm. The solution can alleviate the **non-independent and identically distributed (non-IID)** data distribution problem and limited resources for end users.

In vehicular networks, PPEC nodes, such as at RSUs, can also be effectively achieved by combining FL and DRL frameworks. However, the high mobility of vehicles introduces additional challenges to edge caching efficiency and privacy security. To tackle these challenges, Wu et al. [86] designed an asynchronous FL model to evaluate regional content popularity, taking into account vehicle movement speed, RSU coverage, and network channel conditions. They modified the selection of training vehicles and the aggregation function’s weight, assigning different weights to vehicles with varying dwell times and channel conditions. They proposed a joint content placement strategy based on dueling DRL to overcome the caching efficiency degradation caused by high vehicle mobility. This strategy further reduces content transmission delay while ensuring user data privacy and RSU joint caching efficiency in edge vehicle computing scenarios. Li et al. [39] tackled the privacy and long-term training delay issues in high-precision map caching in intelligent connected vehicles by formulating a framework called *F-DRL* (Federated Deep Reinforcement Learning). F-DRL is an MDP-based edge cooperative caching technique in which Dueling-DQN (Dueling-Deep-Q-Network) is employed to optimize the adaptive edge caching scheme with an improved FL approach to preserve the privacy of the intelligent connected vehicle. For FL, resource provision and member vehicle selection are made using joint optimization to minimize the delays in training and load in the EC.

## 6.2 Combining FL with Other Methods

Other than requiring sensitive data to train machine learning models, the edge caching system may also need private information to make edge caching decisions. Therefore, some works [13, 60, 79, 120] have introduced additional privacy protection methods into the FL framework to enhance data privacy during the model training process. Zheng et al. [120] proposed a privacy-preserving FL model to predict popularity in an unsupervised manner. The prediction method introduces two concepts: local and global popularity, considering both efficiency and privacy. Local popularity can be evaluated by historical information by the LSTM model on users. In contrast, global popularity can only be predicted by the information at the current moment, which will be erased immediately at the next moment. FL is applied to perform offline training and online popularity evaluation with distributed information to avoid exposing privacy. Wang and Deng [79] proposed a private FL-based caching scheme, which utilizes an FL framework and a pseudo rating matrix to collect statistical characteristics of user groups. With this distorted information, the server can predict the popularity of content and make caching decisions. The scheme also protects the privacy of

individual users from being accessed by servers and other users. Saputra et al. [64] introduced the HE method into the FL framework to protect the privacy of MUs with constrained computing resources. The scheme allows MUs to upload encrypted training data to ESs, which can perform additional training processes. The portions of the encrypted decision problem are modeled as a multi-objective profit maximization problem considering both privacy and training costs. The optimization problem is proved to be a concave function that can be solved by the interior point method. At the same time, the training data cached at the EN or the cloud node is HE based on the Brakerski/Fan-Vercauteren, or BFV, method.

### 6.3 Combining Noise-Based FL with Other Methods

More parameters or gradients (representing knowledge extracted from user-related data) can expose user privacy because attackers can probably infer and restore user information from exposed model information. In addition, model parameters may have a huge economic value, and directly uploading model parameters will compromise the self-interest of model owners. Therefore, there are works dedicated to upgrading the FL framework by adding noises [31, 46, 52, 101] or other interference [11, 16, 80] to model information prior to exposure.

The FL framework has further employed DP-based noise to safeguard the parameters or gradients in previous works. Lu et al. [46] designed a differentially private asynchronous FL scheme to share resources in vehicular networks. The proposed scheme uses LDP to perturb the local model parameters with noise drawn from the Gaussian distribution. Moreover, a distributed random update method is used to preserve the privacy of the global ML model during the update process. Yu et al. [101] proposed an FL framework based on privacy protection so that the user dataset is always kept locally. Further, the LDP mechanism is added while exchanging model parameters for aggregation to protect user privacy. In addition, this work proposes a hierarchical joint caching mechanism to combine the characteristics of local caching and global caching. A weighted aggregation method is used to solve the data imbalance problem. Jiang et al. [31] developed a privacy-preserving FL framework for industrial data processing. This framework works by compressing adaptive gradients in the first place during model training at the edge terminal. Afterward, hybrid DP is applied to optimize the FL framework, and the privacy-preserved gradients are transferred in the industrial environment.

Furthermore, some efforts try to enhance privacy protection in edge caching systems by integrating the Generative Adversarial Network technique with FL. Wang et al. [80] combined FL and WGAN (Wasserstein Generative Adversarial Network) to improve further the efficiency of model training and accuracy of the popularity prediction model. With the fake data generated by WGAN, the privacy of users' real preferences can be enhanced. Besides, gradient clipping and model parameter restriction are applied at the training time to protect model privacy and security.

Privacy preservation in vehicular edge computing is demanded since new attack types are developed continuously. To cope with the situation, Chen et al. [11] proposed a novel edge computing approach that utilizes unmanned aerial vehicle swarms as edge computing nodes to aggregate model parameters and caches the model parameters, thereby reducing the communication cost of the core network and protecting users' dataset. To enhance the security and privacy protection of cached model parameters, the authors designed a comprehensive protocol for model aggregation, storage, and transmission, which can effectively prevent potential security threats, such as poisoning attacks, man-in-the-middle attacks, and eavesdropping attacks. Meanwhile, to defend against pollution attacks, the cosine similarity between local parameters and its edge aggregation parameters is calculated to exclude parameters uploaded by malicious nodes. Then, parameters are re-aggregated, and the aggregated parameters are sent to the cloud servers for the final process. A

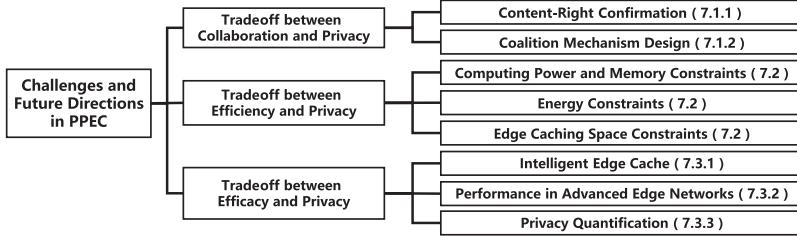


Fig. 7. An outline for challenges and open issues section.

Schnorr signature is also added before uploading the aggregated model parameters to ensure the reliability of the parameters.

In the IoT realm, edge computing architectures can expedite data processing, whereas edge caching can accelerate file delivery speeds for IoT devices. To ensure the reliability and privacy of data in IoT networks, Cui et al. [16] proposed a blockchain system comprising four contracts to predict content popularity, cache, and deliver sensitive content. Meanwhile, to improve the security and throughput of the system, the Proof of Stake (PoS) consensus mechanism based on reputation is modified and applied to reach consensus more efficiently. Besides, the FL algorithm based on compressed gradients is used to protect the privacy information of ESs and reduce communication overhead. The  $K$ -means algorithm filters important gradients that must be uploaded accurately. These gradients are then quantified using a clustering-based quantization algorithm to reduce the amount of data uploaded. Meanwhile, an averaged gradient value is uploaded to the server for other gradients with a small value. Blockchain technique is also used to verify uploaded data. Recently, the Internet of Medical Things has become popular. However, it is also prone to privacy threats like other edge computing based approaches. To tackle these challenges for Internet of Medical Things based big data analytics, Nair et al. [52] proposed an edge computing based FL scheme called *Fed\_select*, which ensures privacy and provides load reduction at the central FL server by introducing an ES. To ensure privacy, *Fed\_select* performs user anonymity at the ES by employing hybrid encryption techniques with client and attribute selection performed at the ES. Moreover, DP with Laplace noise is applied to the shared gradients to make them private during transfer.

Due to limited space, we briefly present an overview of additional solutions for safeguarding knowledge privacy in edge caching systems in Table 5 in the appendix, which covers methods not fully discussed in Section 6.

## 7 Open Challenges and Future Research Directions

In this section, we discuss open challenges and future research directions worth exploring in PPEC. As shown in Figure 7, we elaborate on the challenges and open issues from three major perspectives in PPEC: collaboration, efficiency, and efficacy.

### 7.1 Tradeoff between Collaboration and Privacy in PPEC

Due to the large scale of network applications, it is common to deploy multiple ESs for collaboratively caching content. To enable collaborations between ESs, critical information such as cached content or other private information will be exchanged between ESs, which can expose user privacy and raise privacy concerns. We outline two open privacy concerns when multiple ESs share sensitive information.

**7.1.1 Content-Right Confirmation.** Digital content can be easily copied and distributed, which is a double-edged sword making content-right confirmation difficult. For instance, when social

media content cached on a particular ES is accessed by other ESs, the ES completely loses control of cached social media content because other ESs can easily copy and redistribute this social media content [7, 17, 104]. Content-right confirmation is essential for content owners to maintain availability and accountability when using edge caching services to preserve content privacy. On the one hand, with content-right confirmation, it is easy to determine content ownership. Privacy strategies can be implemented to ensure that only authorized parties can access the content cached in an ES. On the other hand, content-right confirmation is the basis of content accountability. With content-right confirmation, the content right can be authenticated when the content is used once, with the right changed accordingly.

**Challenges.** However, realizing content-right confirmation in edge caching systems is non-trivial due to several challenges. First, the distributed nature of edge caching systems makes it difficult to maintain a centralized and trusted authority for content ownership verification. Second, edge devices' dynamic and heterogeneous nature introduces complexity in implementing content-right confirmation mechanisms, which must be scalable and adaptable to different devices. Furthermore, the use of encryption and privacy-enhancing technologies in edge caching systems further complicates content-right confirmation. Although these technologies are essential for protecting the privacy of cached content, they may also prevent content owners from verifying the use of their content in the cache.

**Future Directions.** The challenges of realizing content-right confirmation in edge caching systems call for future work in several directions. First, new verification mechanisms are needed that can handle the distributed nature, dynamics, and heterogeneity of edge devices. To overcome these challenges, several approaches have been proposed in the literature, such as blockchain-based solutions [38, 77]. However, these mechanisms are short in scalability and the ability to adapt to different edge devices. Second, privacy-preserving verification methods that can coexist with encryption and other privacy-enhancing techniques should be explored. One possible solution is to leverage secure multi-party computation [5] to enable verification while preserving the privacy of cached content. Finally, standardization efforts are needed to ensure interoperability between different edge caching systems and CPs. For example, the trust management mechanisms [94, 121] are proposed to enable content-right confirmation in ISP and D2D edge caching. Thus, promoting the adoption of content-right confirmation mechanisms and facilitating the collaboration between different stakeholders in different edge caching scenarios need to be further discussed.

**7.1.2 Coalition Mechanism Design.** Collaborative edge caching is essential for enhancing QoS. However, many ECs are deployed on leased nodes provided by profit-oriented third-party providers, which are often decentralized [17], unreliable [37], or self-interested [17]. For instance, edge caching routers can be unreliable [17, 37] in a CDN, whereas RSUs and vehicles can be semi-trust [30, 32, 61, 113] or self-interested [14, 19, 34] in the edge IoV caching network. Similarly, in social media networks, ESs can be self-interested [93, 94]. To enable privacy-preserving applications and technology cooperation among edge caching systems, coalition mechanisms are required. These mechanisms involve designing an incentive and allocation model that encourages participants in edge caching systems to join the coalition and maximize their benefits through a reasonable selection. Additionally, punishment mechanisms should be considered when there are untrustworthy or dishonest nodes in the system. Hence, incentive and allocation mechanisms can be designed in a thoughtful manner to foster participation and cooperation among edge caching systems.

**Challenges.** However, designing coalition mechanisms for PPEC is complicated because it is necessary to balance several conflicting objectives. On the one hand, the mechanisms should encourage participants to contribute their resources to the coalition, ensuring that the costs and

benefits of participants are distributed fairly [5]. On the other hand, they must incentivize participants to prioritize the interests of the coalition over their individual interests and punish illegal strategies [93], which is challenging when participants are profit oriented with conflicting goals [94].

**Future Directions.** Game theory is a powerful mathematical framework for investigating decision-making processes, as well as interactions among rational individuals or entities in coalition mechanisms. Its applications can optimize edge caching, capturing interactions among CPs, network operators, and end users. Various game-theoretic models, such as the non-cooperative game [71], Stackelberg game [10, 93, 94], coalition games, and potential games [10], can analyze the interaction among participants in edge caching systems. For instance, in a Stackelberg game, one player acts as the leader while the others follow. In the context of edge caching, the CP can be modeled as the leader while network caching operators are the followers [93]. Similarly, the EC can act as the leader, followed by end users [94]. Nevertheless, game theory based approaches with complete information [93, 94] can not be directly applied in privacy-preserving scenarios since the information is likely incomplete to players in edge caching systems. Besides, it is impractical to assume that every player is benign at an open-access edge network. There may exist semi-honest and even malicious players. Therefore, it is necessary to conduct further investigations into the coalition mechanisms when analyzing the complex interactions between different kinds of players in collaborative PPEC.

## 7.2 Limited Capacity for Running Privacy-Enhancing Caching Algorithms

**Challenges.** Edge devices are becoming increasingly crucial in edge caching networks. However, these devices are typically limited in processing power, memory, caching space, and energy capacity, which present challenges for running privacy-enhancing algorithms. First, limited computing power and memory pose a significant challenge on implementing complex privacy-preserving algorithms on edge devices [54]. To address this challenge, the development of lightweight privacy-preserving algorithms is desired to protect user privacy without compromising caching performance. Lightweight HE [14], identity authentication [75, 95, 96, 113], and DP [2, 71] are prospective approaches that can reconcile privacy protection and computational efficiency. Additionally, it is vital to ensure that the developed algorithms are robust and productive, meeting the needs of edge devices. Second, with regard to energy constraints, edge devices such as autonomous vehicles and smartphones are often battery powered [48], which can limit the ability of caching [64] and communication [16], and hence lower the performance of privacy-preserving algorithms [62, 64]. To address this challenge, energy-efficient caching management techniques should be developed to minimize the energy consumption of PPEC approaches. Techniques such as data compression [16] and optimization models [62] can be adopted to reduce the consumption of caching and communication to minimize energy usage. For instance, an energy-aware client selection and communication method for FL was proposed in the work of Qiao et al. [62] that reduced energy consumption by up to 50% compared to traditional FL methods when protecting the privacy of data sources. Third, cache space is another significant constraint for edge caching systems due to the limited storage capacity compared to the vast amount of content that can be cached. However, research has shown that only a small fraction of content is popular, whereas the majority of users concentrate their access on popular content, implying a long tail distribution of content popularity [47, 81]. Therefore, it is crucial to determine which content should be cached based on popularity and user preferences while considering privacy concerns. PPEC approaches need to tradeoff between privacy protection and caching performance.

**Future Directions.** Implementing privacy-preserving algorithms can complicate the system, which adversely impacts caching performance [42]. Conversely, simplifying the system may

increase the risk of privacy breaches. Challenges associated with PPEC include balancing the complexity of protection algorithms with caching performance [95, 96, 124] or limited resources [16, 71, 75], and ensuring user privacy while enabling efficient content distribution [14, 104]. To sum up, an in-depth understanding of the limitations of edge devices and designing practical and feasible solutions are vital in enhancing PPEC.

### 7.3 Tradeoff between Efficacy and Privacy in PPEC

**7.3.1 Privacy-Enhancing Intelligent EC.** Machine learning based methods have become a powerful tool for optimizing edge caching performance and developing intelligent caching algorithms [45, 62]. However, there has been some controversy regarding privacy violations in edge caching. Privacy concerns arise when edge caching providers analyze and manage content in their cache since the storage spaces of the EC are limited, and the content scale is growing rapidly [81]. To provide intelligent edge caching, providers may be curious about the content stored in their cache (e.g., popular content [7, 14, 17]) and the confidential information about consumers (e.g., request record [55, 85, 104], identifiable information [53, 95, 96, 113]). Providers may use monitoring and inference attacks to compromise consumers' privacy to improve caching efficiency and gain economic benefits. Additionally, the rise of generative AI applications can further complicate this landscape. Generative AI can be leveraged to design more sophisticated caching algorithms that model user behavior and more accurately anticipate content popularity [74]. Furthermore, by caching pre-trained foundation models at the edge network [92], various multimedia enhancement techniques, such as super-resolution, can be deployed to effectively improve user QoE and reduce transport delays. However, these improvements may also come with increased privacy risks, as these models might require more granular user data for fine-tuning at ESs, potentially exposing sensitive information. Therefore, developing effective privacy-preserving mechanisms in intelligent edge caching algorithms is crucial to address these problems.

**Challenges.** However, the open-edge network provides an ideal entrance or interface for attackers to obtain private data or knowledge from machine learning methods designed for edge caching systems. Therefore, reconciling privacy and efficiency in intelligent caching methods at the open-edge network is challenging for several reasons. First, the diversity of user requirements and content in machine learning based edge networks can be more significant than traditional caching systems [122], which makes it difficult to apply traditional privacy-enhancing mechanisms directly. Second, the edge network is usually open, and multi-access [94], implying that it is difficult to control the access of the cache so as to preserve privacy. Finally, the semi-trust or unreliable third-party caching service providers exacerbate the challenge to the design of protection method [7, 45, 62, 95, 96]. For instance, the introduction of generative AI models could intensify these challenges, as these models may require ongoing access to real-time data for training content-specific generative models [92], thereby creating new vectors for privacy breaches. Therefore, designing intelligent caching methods that can well balance privacy and efficiency is challenging.

**Future Directions.** The FL framework is one of the essential methods to preserve private data in the machine learning process [83, 84]. Based on the FL framework, different parties may distributedly predict the critical information, such as content population [45, 62, 120] or user location migratory pattern [86], for intelligent caching decisions at the edge network. However, some parties may be dishonest and malicious. In particular, malicious users in FL may bring poisoned data to affect the overall computing of the global model. For example, dishonest parties may back-infer their partners' model by collecting their gradients to infer private information [16, 101]. These attacks can lead to the disclosure of critical privacy information or destroy caching performance. Additionally, some adversaries even deliberately provide incorrect model

parameters during lateral FL to disrupt the overall computation and impact model performance [80]. The security of private computing in the FL framework is also a challenging open topic for edge caching. Furthermore, generative AI could introduce new methods for obfuscating or anonymizing data before it is cached [80], thereby adding an additional layer of privacy protection. Similarly, with FL or other distributed learning frameworks, a lightweight generative model can be fine-tuned or trained with local data, stored, and run on edge devices, providing personalized and customized AIGC services in real time while maintaining user privacy [91].

**7.3.2 Reconcile Efficacy and Privacy in Advanced Edge Networks.** Traditional edge caching algorithms often rely on pre-determined and suboptimal caching policies, which may not be effective in dynamic network environments where conditions [107] and user behavior [47] can change rapidly over time. In modern network systems, resource access patterns are highly dynamic and complex [114, 119]. Additionally, factors such as user interests, geographical locations, and IoT device connectivity in edge scenarios are also highly dynamic, making one-time trained models less adaptable to these evolving conditions [12]. Some existing works propose high-cost training of machine learning models, which may not be feasible for resource-constrained edge or terminal devices [15]. Furthermore, advanced network infrastructures, such as NDNs and ICNs, are becoming more prevalent in edge networks, introducing new challenges for privacy preservation.

**Challenges.** Several open challenges exist in designing privacy-preserving algorithms for advanced network systems. One challenge is to balance the privacy protection strength and the accuracy of the model prediction. The decision-making process in online scenarios is already highly challenging, and the introduction of privacy protection methods, such as noise perturbation, can further compromise the algorithm's performance or even make it unusable. Another challenge is to develop privacy-preserving algorithms that are computationally efficient and can be easily deployed in dynamic network environments. Moreover, advanced network architectures inherently focus on data rather than specific endpoints, leading to new vulnerabilities. For instance, in the NDN and ICN, content is named and cached throughout the network, which could increase privacy risks if sensitive data is cached without proper safeguards [17, 71, 95]. The ability to cache and retrieve data based on content names rather than IP addresses can expose more granular user preferences and access patterns, making it easier for adversaries to infer sensitive information. Additionally, these architectures may complicate the implementation of privacy-preserving caching strategies, as they require more sophisticated mechanisms to control data access and ensure data integrity across distributed caches.

**Future Directions.** Online learning algorithms, such as reinforcement learning [87] and continuous learning, have become increasingly popular for solving complex problems in various fields, including edge caching in dynamic network environments [12, 15, 35, 107]. However, online learning algorithms can also pose a risk to user privacy when collecting and processing sensitive user data. Therefore, how to safely use the latest historical information to make efficient online caching decisions is a problem worthy of discussion. There are little efforts to address the privacy concerns associated with online learning algorithms. The FL may be a possible framework to allow multiple parties to process data jointly without revealing their raw datasets in dynamic scenarios [35]. Additionally, DP techniques can be used to add random noises to the data in advanced caching systems to obscure individual information [123].

**7.3.3 Privacy Quantification for PPEC.** Privacy quantification is a critical aspect of PPEC systems, as it allows for the measurement and assessment of privacy protection levels provided by these systems [2, 71, 85, 97]. However, most current work on privacy-enhanced intelligent edge caching lacks specific privacy metrics. Rather than developing clear and effective privacy metrics, researchers often combine existing privacy protection schemes and claim that their works can

protect privacy. Unfortunately, without clear quantification of the privacy protection effect, it fails to identify weaknesses for improving privacy protection [2, 97].

**Challenges.** One of the primary challenges in privacy quantification is developing an accurate and consistent metric for measuring privacy protection levels. It is a complicated task to propose a universal method to measure different types of data leakage in the EC. Therefore, an appropriate privacy quantification method with a formalized definition should be established to guide the design of PPEC systems. For instance, in intelligent caching algorithms based on reinforcement learning models, a good privacy exposure quantitative index can guide the model's reward design and help the agent make better caching decisions [93]. Various privacy metrics have been proposed in the literature, such as the information-theoretic converse bound [97] and the size of the anonymity set [85]. However, each metric has its limitations and may not be suitable for general PPEC systems. Evaluating the privacy protection degree in dynamic network environments is another challenge in PPEC. Edge caching systems operate in a constantly changing environment, and various factors can impact privacy protection levels, which makes it challenging to determine an accurate and consistent privacy metric that can be applied in a dynamic environment.

**Future Directions.** To address these challenges, researchers can explore the use of online algorithms [41, 58] to predict privacy protection levels in real time based on network traffic patterns and user behavior. This approach can help to dynamically adjust privacy protection levels in response to changes in the network environment and improve the effectiveness of PPEC. Future work for designing privacy metrics (similar to the privacy budget in DP [2, 71] and mutual information in information theory [71]) is desired. An innovative definition of privacy measurement applicable in intelligent edge caching scenarios [116] should be designed to guide PPEC. Finally, exploring the integration of blockchain technology in PPEC could offer new avenues for ensuring data integrity and transparency [16, 61]. Blockchain can be used to create an immutable record of data access and modifications, thus providing a reliable audit trail that enhances trust and accountability in edge caching systems.

## 8 Conclusion

Edge caching has shown significant potential for improving network performance and resource utilization, but privacy concerns must be considered when deploying ECs. This article analyzed and summarized the most prominent privacy issues in edge caching systems from a *sensitive information* perspective, based on which a comprehensive classification has been proposed. The recent countermeasures for alleviating the exposed threats of different private information were retrospectively reviewed. The article concluded with lessons learned and highlights open challenges for future research in PPEC. Further investigations are needed to ensure the privacy and performance of edge caching while also reconciling the tradeoff between privacy protection and caching performance optimization.

## References

- [1] Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'16)*. ACM, 308–318.
- [2] Gergely Acs, Mauro Conti, Paolo Gasti, Cesar Ghali, Gene Tsudik, and Christopher A. Wood. 2019. Privacy-aware caching in information-centric networking. *IEEE Transactions on Dependable and Secure Computing* 16, 2 (March 2019), 313–328.
- [3] Shahriyar Amini, Janne Lindqvist, Jason I. Hong, Maladau Mou, Rahul Raheja, Jialiu Lin, Norman Sadeh, and Eran Toch.. 2011. Caché: Caching location-enhanced content to improve user privacy. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services (MobiSys'11)*. ACM, 197–209.

- [4] Davide Andreoletti, Omran Ayoub, Silvia Giordano, Giacomo Verticale, and Massimo Tornatore. 2019. Privacy-preserving caching in ISP networks. In *Proceedings of the IEEE 20th International Conference on High Performance Switching and Routing (HPSR'19)*. IEEE, 1–6.
- [5] Davide Andreoletti, Silvia Giordano, Giacomo Verticale, and Massimo Tornatore. 2018. Discovering the geographic distribution of live videos' users: A privacy-preserving approach. In *Proceedings of the IEEE Global Communications Conference (GLOBECOM'18)*. IEEE, 1–6.
- [6] Davide Andreoletti, Cristina Rottandi, Silvia Giordano, Giacomo Verticale, and Massimo Tornatore. 2019. An open privacy-preserving and scalable protocol for a network-neutrality compliant caching. In *Proceedings of the IEEE International Conference on Communications (ICC'19)*. IEEE, 1–6.
- [7] Andrea Araldo, Gyorgy Dan, and Dario Rossi. 2018. Caching encrypted content via stochastic cache partitioning. *IEEE/ACM Transactions on Networking* 26, 1 (January 2018), 548–561.
- [8] Paolo Bellavista, Luca Foschini, and Alessio Mora. 2021. Decentralised learning in federated deployment environments: A system-level survey. *ACM Computing Surveys* 54, 1 (Februar 2021), Article 15, 38 pages.
- [9] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. 1273–1282.
- [10] Tengfei Cao, Changqiao Xu, Jumping Du, Yawen Li, Han Xiao, and Changhui Gong. 2020. Reliable and efficient multimedia service optimization for edge computing-based 5G networks: Game theoretic approaches. *IEEE Transactions on Network and Service Management* 17, 3 (September 2020), 1610–1625.
- [11] Qi Chen, Bing Chen, Feng Hu, and Jiale Zhang. 2022. Edge-based protection against malicious poisoning for distributed federated learning. In *Proceedings of the IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD'22)*. IEEE, 459–464.
- [12] Zhiqi Chen, Sheng Zhang, Zhi Ma, Shuai Zhang, Zhuzhong Qian, and Mingjun Xiao. 2022. An online approach for DNN model caching and processor allocation in edge computing. In *Proceedings of the IEEE/ACM 30th International Symposium on Quality of Service (IWQoS'22)*. IEEE, 1–10.
- [13] Runze Cheng, Yao Sun, Yijing Liu, Le Xia, Sanshan Sun, and Muhammad Ali Imran. 2021. A privacy-preserved D2D caching scheme underpinned by blockchain-enabled federated learning. In *Proceedings of the IEEE Global Communications Conference (GLOBECOM'21)*. IEEE, 1–6.
- [14] Jie Cui, Lu Wei, Hong Zhong, Jing Zhang, Xan Wu, and Lu Liu. 2020. Edge computing in VANETs—An efficient and privacy-preserving cooperative downloading scheme. *IEEE Journal on Selected Areas in Communications* 38, 6 (April 2020), 1191–1204.
- [15] Laizhong Cui, Erchao Ni, Yipeng Zhou, Zhi Wang, Lei Zhang, and Jiangchuan Liu. 2023. Towards real-time video caching at edge servers: A cost-aware deep Q-learning solution. *IEEE Transactions on Multimedia* 25 (November 2023), 302–314.
- [16] Laizhong Cui, Xiaoxin Su, Zhongxing Ming, Ziteng Chen, Shu Yang, and Yipeng Zhou. 2022. CREAT: Blockchain-assisted compression algorithm of federated learning for content caching in edge computing. *IEEE Internet of Things Journal* 9, 16 (August 2022), 14151–14161.
- [17] Shujie Cui, Muhammad Rizwan Asghar, and Giovanni Russello. 2020. Multi-CDN: Towards privacy in content delivery networks. *IEEE Transactions on Dependable and Secure Computing* 17, 5 (September 2020), 984–999.
- [18] Yuanbo Cui, Fei Gao, Wenmin Li, Yijie Shi, Hua Zhang, Qiaoyan Wen, and Emmanouil Panaousis. 2020. Cache-based privacy preserving solution for location and content protection in location-based services. *Sensors* 20, 16 (August 2020), 4651.
- [19] Yueyue Dai, Du Xu, Ke Zhang, Sabita Maharjan, and Yan Zhang. 2020. Deep reinforcement learning and permissioned blockchain for content caching in vehicular edge computing and networks. *IEEE Transactions on Vehicular Technology* 69, 4 (April 2020), 4312–4324.
- [20] Subhankar Dhar and Upkar Varshney. 2011. Challenges and business models for mobile location-based services and advertising. *Communications of the ACM* 54, 5 (May 2011), 121–129.
- [21] Jeroen Famaey, Frédéric Iterbeke, Tim Wauters, and Filip De Turck. 2013. Towards a predictive cache replacement strategy for multimedia content. *Journal of Network and Computer Applications* 36, 1 (January 2013), 219–227.
- [22] Haojie Feng, Yanxiang Jiang, Dusit Niyato, Fu-Chun Zheng, and Xiaohu You. 2019. Content popularity prediction via deep learning in cache-enabled fog radio access networks. In *Proceedings of the IEEE Global Communications Conference (GLOBECOM'19)*. IEEE, 1–6.
- [23] Lidia Fotia, Flávia Delicato, and Giancarlo Fortino. 2023. Trust in edge-based Internet of Things architectures: State of the art and research challenges. *ACM Computing Surveys* 55, 9 (January 2023), Article 182, 34 pages.
- [24] Omer Gil. 2017. *Web Cache Deception Attack*. Technical Report. Black Hat USA.
- [25] Yi-Ming Gu, Guang-Wei Bai, Hang Shen, and Yu-Jia Hu. 2019. Pre-cache based privacy protection mechanism in continuous LBS queries. *Computer Science* 46, 5 (May 2019), 122–128.

- [26] Jianxiong Guo, Xingjian Ding, Tian Wang, and Weijia Jia. 2022. Combinatorial resources auction in decentralized edge-thing systems using blockchain and differential privacy. *Information Sciences* 607 (August 2022), 211–229.
- [27] Seyedeh Bahereh Hassanpour, Abolfazl Diyanat, Ahmad Khonsari, Seyed Pooya Shariatiapanahi, and Aresh Dadlani. 2021. Context-aware privacy preservation in network caching: an information theoretic approach. *IEEE Communications Letters* 25, 1 (January 2021), 54–58.
- [28] Seyedeh Bahereh Hassanpour, Ahmad Khonsari, Masoumeh Moradian, and Seyed Pooya Shariatiapanahi. 2023. Privacy-preserving edge caching: A probabilistic approach. *Computer Networks* 226 (May 2023), 109654.
- [29] Xiuli He, Ashutosh Prasad, Suresh P. Sethi, and Genaro J. Gutierrez. 2007. A survey of Stackelberg differential game models in supply and marketing channels. *Journal of Systems Science and Systems Engineering* 16, 4 (November 2007), 385–413.
- [30] Long Hu, Yongfeng Qian, Min Chen, M. Shamim Hossain, and Ghulam Muhammad. 2018. Proactive cache-based location privacy preserving for vehicle networks. *IEEE Wireless Communications* 25, 6 (December 2018), 77–83.
- [31] Bin Jiang, Jianjiang Li, Huihui Wang, and Houbing Song. 2023. Privacy-preserving federated learning for industrial edge computing via hybrid differential privacy and adaptive compression. *IEEE Transactions on Industrial Informatics* 19, 2 (February 2023), 1136–1144.
- [32] Shunrong Jiang, Jianqing Liu, Longxia Huang, Haiqin Wu, and Yong Zhou. 2020. Vehicular edge computing meets cache: An access control scheme for content delivery. In *Proceedings of the IEEE International Conference on Communications (ICC'20)*. IEEE, 1–6.
- [33] Wei Jiang, Gang Feng, and Shuang Qin. 2017. Optimal cooperative content caching and delivery policy for heterogeneous cellular networks. *IEEE Transactions on Mobile Computing* 16, 5 (May 2017), 1382–1393.
- [34] Qinglei Kong, Rongxing Lu, Maode Ma, and Haiyong Bao. 2019. A privacy-preserving and verifiable querying scheme in vehicular fog data dissemination. *IEEE Transactions on Vehicular Technology* 68, 2 (February 2019), 1877–1887.
- [35] Krishnendu S. Tharakan, B. N. Bharath, Navneet Garg, Vimal Bhatia, and Tharmalingam Ratnarajah. 2022. Learning to cache: Federated caching in a cellular network with correlated demands. *IEEE Transactions on Communications* 70, 3 (March 2022), 1653–1665.
- [36] Siddhartha Kumar, Alexandre I. Graell Amat, Eirik Rosnes, and Linda Senigagliesi. 2019. Private information retrieval from a cellular network with caching at the edge. *IEEE Transactions on Communications* 67, 7 (July 2019), 4900–4912.
- [37] Jeremie Leguay, Georgios S. Paschos, Elizabeth A. Quaglia, and Ben Smyth. 2017. CryptoCache: Network caching with confidentiality. In *Proceedings of the IEEE International Conference on Communications (ICC'17)*. IEEE, 1–6.
- [38] Kai Lei, Junjie Fang, Qichao Zhang, Junjun Lou, Maoyu Du, Jiyue Huang, Jianping Wang, and Kuai Xu. 2020. Blockchain-based cache poisoning security protection and privacy-aware access control in NDN vehicular edge computing networks. *Journal of Grid Computing* 18, 4 (August 2020), 593–613.
- [39] Chunlin Li, Yong Zhang, and Youlong Luo. 2023. A federated learning-based edge caching approach for mobile edge computing-enabled intelligent connected vehicles. *IEEE Transactions on Intelligent Transportation Systems* 24, 3 (November 2023), 3360–3369.
- [40] Ruibin Li, Yiwei Zhao, Chenyang Wang, Xiaofei Weng, Victor C. M. Leung, and Xiuhua Li. 2020. Edge caching replacement optimization for D2D wireless networks via weighted distributed DQN. In *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC'20)*. IEEE, 1–6.
- [41] Weiting Li, Liyao Xiang, Zhou Zhou, and Feng Peng. 2021. Privacy budgeting for growing machine learning datasets. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM'21)*. IEEE, 1–10.
- [42] Jie Liang and Yinlong Liu. 2019. A cache privacy protection strategy based on content privacy and user security classification in CCN. In *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC'19)*. IEEE, 1–6.
- [43] Dong Liu, Binqiang Chen, Chenyang Yang, and Andreas F. Molisch. 2016. Caching at the wireless edge: Design aspects, challenges, and future directions. *IEEE Communications Magazine* 54, 9 (September 2016), 22–28.
- [44] Jiadi Liu, Songtao Guo, Yawei Shi, Liang Feng, and Cong Wang. 2020. Decentralized caching framework toward edge network based on blockchain. *IEEE Internet of Things Journal* 7, 9 (June 2020), 9158–9174.
- [45] Shengheng Liu, Chong Zheng, Yongming Huang, and Tony Q. S. Quek. 2022. Distributed reinforcement learning for privacy-preserving dynamic edge caching. *IEEE Journal on Selected Areas in Communications* 40, 3 (March 2022), 749–760.
- [46] Yunlong Lu, Xiaohong Huang, Yueyue Dai, Sabita Maharjan, and Yan Zhang. 2020. Differentially private asynchronous federated learning for mobile edge computing in urban informatics. *IEEE Transactions on Industrial Informatics* 16, 3 (March 2020), 2134–2143.

[47] Ge Ma, Zhi Wang, Miao Zhang, Jiahui Ye, Minghua Chen, and Wenwu Zhu. 2017. Understanding performance of edge content caching for mobile video streaming. *IEEE Journal on Selected Areas in Communications* 35, 5 (May2017), 1076–1089.

[48] Yuyi Mao, Changsheng You, Jun Zhang, Kaibin Huang, and Khaled B. Letaief. 2017. A survey on mobile edge computing: The communication perspective. *IEEE Communications Surveys and Tutorials* 19, 4 (October2017), 2322–2358.

[49] Seyed Ali Mirheidari, Sajjad Arshad, Kaan Onarlioglu, Bruno Crispo, Engin Kirda, and William Robertson. 2020. Cached and confused: Web cache deception in the wild. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security'20)*. 665–682.

[50] Seyed Ali Mirheidari, Matteo Golinelli, Kaan Onarlioglu, Engin Kirda, and Bruno Crispo. 2022. Web cache deception escalates! In *Proceedings of the 31st USENIX Security Symposium (USENIX Security'22)*. 179–196.

[51] Sabrina Muller, Onur Atan, Mihaela van der Schaar, and Anja Klein. 2017. Context-aware proactive content caching with service differentiation in wireless networks. *IEEE Transactions on Wireless Communications* 16, 2 (February2017), 1024–1036.

[52] Akarsh K. Nair, Jayakrushna Sahoo, and Ebin Deni Raj. 2023. Privacy preserving federated learning framework for IoMT based big data analysis using edge computing. *Computer Standards & Interfaces* 86 (January2023), 103720.

[53] Duong Thuy Anh Nguyen, Jiaming Cheng, Duong Tung Nguyen, and Angelia Nedich. 2023. CrowdCache: A decentralized game-theoretic framework for mobile edge content sharing. *arxiv:cs.GT/2304.13246* (2023). <https://arxiv.org/abs/2304.13246>

[54] Jianbing Ni, Kuan Zhang, and Athanasios V. Vasilakos. 2021. Security and privacy for mobile edge caching: Challenges and solutions. *IEEE Wireless Communications* 28, 3 (June2021), 77–83.

[55] Stavros Nikolaou, Robbert Van Renesse, and Nicolas Schiper. 2016. Proactive cache placement on cooperative client caches for online social networks. *IEEE Transactions on Parallel and Distributed Systems* 27, 4 (April2016), 1174–1186.

[56] Nisha Nisha, Iynkaran Natgunanathan, Shang Gao, and Yong Xiang. 2022. A novel privacy protection scheme for location-based services using collaborative caching. *Computer Networks* 213 (August2022), 109107.

[57] Yazan Otoum, Navya Gottumukkala, Neeraj Kumar, and Amiya Nayak. 2024. Machine learning in metaverse security: Current solutions and future challenges. *ACM Computing Surveys* 56, 8 (April 2024), Article 215, 36 pages.

[58] Xiaoyi Pang, Zhibo Wang, Jingxin Li, Ruiting Zhou, Ju Ren, and Zhetao Li. 2022. Towards online privacy-preserving computation offloading in mobile edge computing. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM'22)*. IEEE, 1179–1188.

[59] Yuwen Pu, Ying Wang, Feihong Yang, Jin Luo, Chunqiang Hu, and Haibo Hu. 2019. An efficient and recoverable data sharing mechanism for edge storage. In *Proceedings of the International Conference on Wireless Algorithms, Systems, and Applications (WASA'19)*. 247–259.

[60] Kaiqiang Qi and Chenyang Yang. 2020. Popularity prediction with federated learning for proactive caching at wireless edge. In *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC'20)*. IEEE, 1–6.

[61] Yongfeng Qian, Yingying Jiang, Long Hu, M. Shamim Hossain, Mubarak Alrashoud, and Muneer Al-Hammadi. 2020. Blockchain-based privacy-aware content caching in Cognitive Internet of Vehicles. *IEEE Network* 34, 2 (March2020), 46–51.

[62] Dewen Qiao, Songtao Guo, Defang Liu, Saiqin Long, Pengzhan Zhou, and Zhetao Li. 2022. Adaptive federated deep reinforcement learning for proactive content caching in edge computing. *IEEE Transactions on Parallel and Distributed Systems* 33, 12 (December2022), 4767–4782.

[63] Ju Ren, Deyu Zhang, Shiwen He, Yaoxue Zhang, and Tao Li. 2019. A survey on end-edge-cloud orchestrated network computing paradigms: Transparent computing, mobile edge computing, fog computing, and cloudlet. *ACM Computing Surveys* 52, 6 (October 2019), Article 125, 36 pages.

[64] Yuris Mulya Saputra, Diep N. Nguyen, Dinh Thai Hoang, and Eryk Dutkiewicz. 2022. In-network caching and learning optimization for federated learning in mobile edge networks. In *Proceedings of the IEEE International Conference on Communications (ICC'22)*. IEEE, 1653–1658.

[65] Kinza Sarwar, Sira Yongchareon, Jian Yu, and Saeed Ur Rehman. 2021. A survey on privacy preservation in fog-enabled Internet of Things. *ACM Computing Surveys* 55, 1 (November 2021), Article 2, 39 pages.

[66] Reent Schlegel, Siddhartha Kumar, Eirik Rosnes, and Alexandre Graell i Amat. 2022. Privacy-preserving coded mobile edge computing for low-latency distributed inference. *IEEE Journal on Selected Areas in Communications* 40, 3 (March2022), 788–799.

[67] Adnan A. Abi Sen, Fathy B. Eassa, Mohammad Yamin, and Kamal Jambi. 2018. Double cache approach with wireless technology for preserving user privacy. *Wireless Communications and Mobile Computing* 2018 (January2018), 1–11.

[68] M. Zubair Shafiq, Alex X. Liu, and Amir R. Khakpour. 2014. Revisiting caching in content delivery networks. *ACM SIGMETRICS Performance Evaluation Review* 42, 1 (2014), 567–568.

[69] Fang Shi, Lisheng Fan, Xin Liu, Zhenyu Na, and Yanchen Liu. 2018. Probabilistic caching placement in the presence of multiple eavesdroppers. *Wireless Communications and Mobile Computing* 2018 (May2018), 1–10.

[70] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P'17)*. IEEE, 3–18.

[71] Vignesh Sivaraman and Biplab Sikdar. 2021. A defense mechanism against timing attacks on user privacy in ICN. *IEEE/ACM Transactions on Networking* 29, 6 (December2021), 2709–2722.

[72] Splunk Inc.2023. Internet Trends in 2024: Mary Meeker, Stats, & Predictions. Retrieved July 21, 2024 from [https://www.splunk.com/en\\_us/blog/learn/internet-trends.html](https://www.splunk.com/en_us/blog/learn/internet-trends.html)

[73] R. S. Sutton and A. G. Barto. 1998. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks* 9, 5 (1998), 1054.

[74] Shisong Tang, Qing Li, Xiaoteng Ma, Ci Gao, Dingmin Wang, Yong Jiang, Qian Ma, Aoyang Zhang, and Hechang Chen. 2022. Knowledge-based temporal fusion network for interpretable online video popularity prediction. In *Proceedings of the ACM Web Conference (WWW'22)*. ACM, 2879–2887.

[75] Wei Tong, Wenjie Chen, Bingbing Jiang, Fengyuan Xu, Qun Li, and Sheng Zhong. 2023. Privacy-preserving data integrity verification for secure mobile edge storage. *IEEE Transactions on Mobile Computing* 22, 9 (March2023), 5463–5478.

[76] Reza Tourani, Satyajayant Misra, Travis Mick, and Guarav Panwar. 2018. Security, privacy, and access control in information-centric networking: A survey. *IEEE Communications Surveys and Tutorials* 20, 1 (January2018), 556–600.

[77] Thang X. Vu, Symeon Chatzinotas, and Bjorn Ottersten. 2019. Blockchain-based content delivery networks: Content transparency meets user privacy. In *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC'19)*. IEEE, 1–6.

[78] Huanhuan Wang, Xiao Zhang, Youbing Xia, and Xiang Wu. 2022. A differential privacy-preserving deep learning caching framework for heterogeneous communication network systems. *International Journal of Intelligent Systems* 37, 12 (August2022), 11142–11166.

[79] Kailun Wang and Na Deng. 2022. A privacy-protected popularity prediction scheme for content caching based on federated learning. *IEEE Transactions on Vehicular Technology* 71, 9 (June2022), 10191–10196.

[80] Kailun Wang, Na Deng, and Xuanheng Li. 2022. An efficient content popularity prediction of privacy preserving based on federated learning and Wasserstein GAN. *IEEE Internet of Things Journal* 10, 5 (May2022), 3786–3798.

[81] Mu Wang, Changqiao Xu, Xingyan Chen, Hao Hao, Lujie Zhong, and Shui Yu. 2019. Differential privacy oriented distributed online learning for mobile social video prefetching. *IEEE Transactions on Multimedia* 21, 3 (January2019), 636–651.

[82] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. 2017. Locally differentially private protocols for frequency estimation. In *Proceedings of the 26th USENIX Security Symposium*. ACM, 729–745.

[83] Xiaofei Wang, Yiwen Han, Chenyang Wang, Qiyang Zhao, Xu Chen, and Min Chen. 2019. In-Edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning. *IEEE Network* 33, 5 (September2019), 156–165.

[84] Xiaofei Wang, Chenyang Wang, Xiuhua Li, Victor C. M. Leung, and Tarik Taleb. 2020. Federated deep reinforcement learning for Internet of Things with decentralized cooperative edge caching. *IEEE Internet of Things Journal* 7, 10 (October2020), 9441–9455.

[85] Qinghua Wu, Zhenyu Li, Gareth Tyson, Steve Uhlig, Mohamed Ali Kaafar, and Gaogang Xie. 2016. Privacy-aware multipath video caching for content-centric networks. *IEEE Journal on Selected Areas in Communications* 34, 8 (August2016), 2219–2230.

[86] Qiong Wu, Yu Zhao, Qiang Fan, Pingyi Fan, Jiangzhou Wang, and Cui Zhang. 2023. Mobility-aware cooperative caching in vehicular edge computing based on asynchronous federated and deep reinforcement learning. *IEEE Journal on Selected Topics in Signal Processing* 17, 1 (January2023), 66–81.

[87] Liang Xiao, Xiaoyue Wan, Canhuang Dai, Xiaojiang Du, Xiang Chen, and Mohsen Guizani. 2018. Security in mobile edge caching with reinforcement learning. *IEEE Wireless Communications* 25, 3 (June2018), 116–122.

[88] Linchang Xiao, Xianzhi Zhang, Di Wu, Miao Hu, Yipeng Zhou, and Shui Yu. 2024. History-aware privacy budget allocation for model training on evolving data-sharing platforms. *IEEE Transactions on Services Computing*. Preprint.

[89] Jinbo Xiong, Renwan Bi, Mingfeng Zhao, Jingda Guo, and Qing Yang. 2020. Edge-assisted privacy-preserving raw data sharing framework for connected autonomous vehicles. *IEEE Wireless Communications* 27, 3 (2020), 24–30.

[90] Sijie Xiong, Anand D. Sarwate, and Narayan B. Mandayam. 2022. Network traffic shaping for enhancing privacy in IoT systems. *IEEE/ACM Transactions on Networking* 30, 3 (2022), 1162–1177.

[91] Minrui Xu, Hongyang Du, Dusit Niyato, Jiawen Kang, Zehui Xiong, and Shiwen Mao. 2024. Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services. *IEEE Communications Surveys & Tutorials* 26, 2 (2024), 1127–1170.

[92] Minrui Xu, Dusit Niyato, Hongliang Zhang, Jiawen Kang, Zehui Xiong, Shiwen Mao, and Zhu Han. 2023. Joint foundation model caching and inference of generative AI services for edge intelligence. In *Proceedings of the IEEE Global Communications Conference (GLOBECOM'23)*. IEEE, 3548–3553.

[93] Qichao Xu, Zhou Su, Rongxing Lu. 2020. Game theory and reinforcement learning based secure edge caching in mobile social networks. *IEEE Transactions on Information Forensics and Security* 15 (March2020), 3415–3429.

[94] Qichao Xu, Zhou Su, Qinghua Zheng, Minnan Luo, Bo Dong, and Kuan Zhang. 2019. Game theoretical secure caching scheme in multihoming edge computing-enabled heterogeneous networks. *IEEE Internet of Things Journal* 6, 3 (June2019), 4536–4546.

[95] Kaiping Xue, Peixuan He, Xiang Zhang, Qiudong Xia, David S. L. Wei, and Hao Yue. 2019. A secure, efficient, and accountable edge-based access control framework for information centric networks. *IEEE/ACM Transactions on Networking* 27, 3 (June2019), 1220–1233.

[96] Kaiping Xue, Xiang Zhang, Qiudong Xia, David S. L. Wei, Hao Yue, and Feng Wu. 2018. SEAF: A secure, efficient and accountable access control framework for information centric networking. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM'18)*. IEEE, 2213–2221.

[97] Qifa Yan and Daniela Tuninetti. 2021. Fundamental limits of caching for demand privacy against colluding users. *IEEE Journal on Selected Areas in Information Theory* 2, 1 (January2021), 192–207.

[98] Peng Yang, Ning Zhang, Shan Zhang, Li Yu, Junshan Zhang, and Xuemin Shen. 2019. Content popularity prediction towards location-aware mobile edge caching. *IEEE Transactions on Multimedia* 21, 4 (April2019), 915–929.

[99] Qiuwei Yang and Pan Kong. 2016. RuleCache: A mobility pattern based multi-level cache approach for location privacy protection. In *Proceedings of the IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS'16)*. IEEE, 448–455.

[100] Zhengxin Yu, Jia Hu, Geyong Min, Haochuan Lu, Zhiwei Zhao, and Haozhe Wang. 2018. Federated learning based proactive content caching in edge computing. In *Proceedings of the IEEE Global Communications Conference (GLOBECOM'18)*. IEEE, 1–6.

[101] Zhengxin Yu, Jia Hu, Geyong Min, Zi Wang, Wang Miao, and Shancang Li. 2022. Privacy-preserving federated deep learning for cooperative hierarchical caching in fog computing. *IEEE Internet of Things Journal* 9, 22 (May2022), 22246–22255.

[102] Zhengxin Yu, Jia Hu, Geyong Min, Han Xu, and Jed Mills. 2020. Proactive content caching for Internet-of-Vehicles based on peer-to-peer federated learning. In *Proceedings of the IEEE 26th International Conference on Parallel and Distributed Systems (ICPADS'20)*. IEEE, 601–608.

[103] Zhengxin Yu, Jia Hu, Geyong Min, Zhiwei Zhao, Wang Miao, and M. Shamim Hossain. 2021. Mobility-aware proactive edge caching for connected vehicles using federated learning. *IEEE Transactions on Intelligent Transportation Systems* 22, 8 (August2021), 5341–5351.

[104] Xingliang Yuan, Xinyu Wang, Jinfang Wang, Yilei Chu, Cong Wang, and Jianping Wang. 2016. Enabling secure and efficient video delivery through encrypted in-network caching. *IEEE Journal on Selected Areas in Communications* 34, 8 (August2016), 2077–2090.

[105] Yiming Zeng, Yaodong Huang, Ji Liu, and Yuanyuan Yang. 2020. Privacy-preserving distributed edge caching for mobile data offloading in 5G networks. In *Proceedings of the IEEE 40th International Conference on Distributed Computing Systems (ICDCS'20)*. IEEE, 541–551.

[106] Yiming Zeng, Yaodong Huang, Zehnhua Liu, Ji Liu, and Yuanyuan Yang. 2021. Privacy-preserving decentralized edge caching in 5G networks. In *Proceedings of the IEEE 14th International Conference on Cloud Computing (CLOUD'21)*. IEEE, 189–199.

[107] Chi Zhang, Haisheng Tan, Guopeng Li, Zhenhua Han, Shaofeng H.-C. Jiang, and Xiang-Yang Li. 2022. Online file caching in latency-sensitive systems with delayed hits and bypassing. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM'22)*. IEEE, 1059–1068.

[108] Jie Zhang, Zhihao Qu, Chenxi Chen, and Wang Haozhao. 2021. Edge learning: The enabling technology for distributed big data analytics in the edge. *ACM Computing Surveys* 54, 7 (July 2021), Article 151, 36 pages.

[109] Shiwen Zhang, Biao Hu, Wei Liang, Kuan-Ching Li, and Brij B. Gupta. 2023. A caching-based dual  $K$ -anonymous location privacy-preserving scheme for edge computing. *IEEE Internet of Things Journal* 10, 11 (January2023), 9768–9781.

[110] Shaobo Zhang, Xiong Li, Zhiyuan Tan, Tao Peng, and Guojun Wang. 2019. A caching and spatial  $K$ -anonymity driven privacy enhancement scheme in continuous location-based services. *Future Generation Computer Systems* 94 (May2019), 40–50.

[111] Xinyue Zhang, Jingyi Wang, Hongning Li, Yuanxiong Guo, Qingpi Pei, and Pan Li. 2018. Data-driven caching with users' local differential privacy in information-centric networks. In *Proceedings of the IEEE Global Communications Conference (GLOBECOM'18)*. IEEE, 1–6.

- [112] Xianzhi Zhang, Linchang Xiao, Yipeng Zhou, Miao Hu, Di Wu, John C. S. Lui, and Quan Z. Sheng. 2023. cRVR: A Stackelberg game approach for joint privacy-aware video requesting and edge caching. *arxiv:cs.NI/2310.12622* (2023). <https://arxiv.org/abs/2310.12622>
- [113] Xiaoyu Zhang, Hong Zhong, Chunyang Fan, Irina Bolodurina, and Jie Cui. 2022. CBACS: A privacy-preserving and efficient cache-based access control scheme for software defined vehicular networks. *IEEE Transactions on Information Forensics and Security* 17 (May2022), 1930–1945.
- [114] Xianzhi Zhang, Yipeng Zhou, Di Wu, Miao Hu, James Xi Zheng, Min Chen, and Song Guo. 2022. Optimizing video caching at the edge: A hybrid multi-point process approach. *IEEE Transactions on Parallel and Distributed Systems* 33, 10 (October2022), 2597–2611.
- [115] Xianzhi Zhang, Yipeng Zhou, Di Wu, Quan Z. Sheng, Miao Hu, and Linchang Xiao. 2024. PPVF: An efficient privacy-preserving online video fetching framework with correlated differential privacy. *arxiv:cs.MM/2408.14735* (2024) <https://arxiv.org/abs/2408.14735>
- [116] Zizhen Zhang, Tengfei Cao, Xiaoying Wang, Han Xiao, and Jianfeng Guan. 2022. VC-PPQ: Privacy-preserving Q-learning based video caching optimization in mobile edge networks. *IEEE Transactions on Network Science and Engineering* 9, 6 (August2022), 4129–4144.
- [117] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. 2020. iDLG: Improved deep leakage from gradients. *arxiv:cs.LG/2001.02610* (2020) <https://arxiv.org/abs/2001.02610>
- [118] Kongyange Zhao, Zhi Zhou, Xu Chen, Ruiting Zhou, Xiaoxi Zhang, Shuai Yu, and Di Wu. 2023. EdgeAdaptor: Online configuration adaption, model selection and resource provisioning for edge DNN inference serving at scale. *IEEE Transactions on Mobile Computing* 22, 10 (2023), 5870–5886.
- [119] Kongyange Zhao, Zhi Zhou, Lei Jiao, Shen Cai, Fei Xu, and Xu Chen. 2024. Taming serverless cold start of cloud model inference with edge computing. *IEEE Transactions on Mobile Computing* 23, 8 (2024), 8111–8128.
- [120] Chong Zheng, Shengheng Liu, Yongming Huang, Wei Zhang, and Luxi Yang. 2022. Unsupervised recurrent federated learning for edge popularity prediction in privacy-preserving mobile-edge computing networks. *IEEE Internet of Things Journal* 9, 23 (2022), 24328–24345.
- [121] Yuqing Zhong, Zhaohua Li, and Liping Liao. 2021. A privacy-preserving caching scheme for device-to-device communications. *Security and Communication Networks* 2021 (January2021), 10958.
- [122] Pan Zhou, Kehao Wang, Jie Xu, and Dapeng Wu. 2019. Differentially-private and trustworthy online social multi-media big data retrieval in edge computing. *IEEE Transactions on Multimedia* 21, 3 (March2019), 539–554.
- [123] Yipeng Zhou, Xuezheng Liu, Yao Fu, Di Wu, Chao Li, and Shui Yu. 2023. Optimizing the numbers of queries and replies in convex federated learning with differential privacy. *IEEE Transactions on Dependable and Secure Computing* 20, 6 (November2023), 4823–4837.
- [124] Pengcheng Zhu, Jun Xu, Jiamin Li, Dongming Wang, and Xiaohu You. 2021. Learning-empowered privacy preservation in beyond 5G edge intelligence networks. *IEEE Wireless Communications* 28, 2 (April2021), 12–18.

Received 24 August 2023; revised 2 October 2024; accepted 23 November 2024