Communication-Efficient Private Information Acquisition: Multicasting via Crowding

Hyowoon Seo, *Member, IEEE*, Kyungrak Son, *Student Member, IEEE*, Sangjun Park, *Student Member, IEEE*, and Wan Choi, *Fellow, IEEE*

***Abstract*—This paper focuses on the way to protect privacy of clients requesting datasets stored in data servers while keeping communication efficiency. To this end, we introduce a novel communication-efficient and privacy protecting framework termed *crowded information acquisition (CIA)*, well suited to a large number of clients scenario. We investigate the CIA under various conditions addressing possible communication scenarios. Contrary to the conventional belief, the results claim that a large number of clients demanding private services can enhance the privacy protection while providing low latency services and generating a small amount of traffic.**

***Index Terms*—Private information retrieval, crowded information acquisition, communication efficiency**

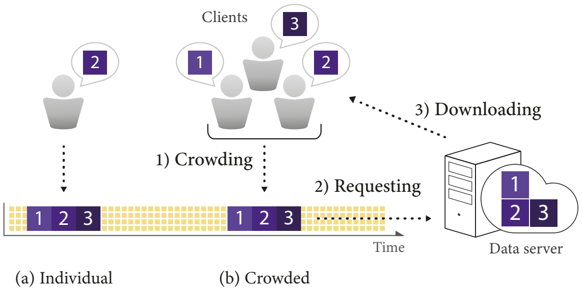


Fig. 1. An overview of the conventional (a) individual information acquisition and the proposed (b) crowded information acquisition.

I. Introduction

Many people today are hyperconnected to the Internet by the help of advanced communication systems and devices. This high connectivity surely offers a great convenience to the people, however, it may also lead to serious and incognizable privacy leakage problems. One possible scenario of a client's privacy disclosure to a server happens when the client sends query for information acquisition, e.g., medical data collection, streaming or downloading videos, using GPS-based maps on vehicles and searching the Internet. Through the query sent by the client, the server may collect private information, such as interests, intentions, etc., whether the client wants it or not.

Unfortunately, in the networks that demand user registration, e.g., cellular networks, V2X networks, private networks, etc, the clients are unable to leverage anonymity-based approaches for protecting their privacy. Alternatively, one of the simplest and effective actions that the client can take is to request undesired dummy information together with the desired information, in order to increase uncertainty of the client's desire in the perspective of the server. The problem of such a simple approach is that due to a large amount of dummy information for increasing the uncertainty, it wastes a large amount of communication resources for getting the desired information and causes traffic problems. To resolve this, a private information retrieval (PIR) method [1], [2] and its variations [3]–[5] suggest the concept of getting coded information from multiple database servers and canceling out the dummy information with help of diversity gain from the multiple servers. Yet, if there are a large number of clients that are trying to acquire the desired information with privacy protective methods, the system will still suffer from serious latency and traffic issues. In this context, designing a *communication-efficient and private information acquisition method* for supporting a large number of clients, without sparing excessive amount of communication resources, is the prime interest of this paper. To this end, we first seek for a novel approach that can effectively reduce the service delay and network traffic, which is very challenging in general. Fortunately, the traditional concept of *hiding in the crowd* helps us to come up with a simple-but-powerful idea: *when a group of clients cooperatively requests the desired information to a data server, the server can hardly figure out the individual interest of the members in the group, and much harder if the number of members in the group is very large.* Here, the idea is that instead of hiding a client in the crowd for identity anonymization, its individual interests are hidden in crowds' interests to protect the individual privacy about its interests, without anonymizing the identity of the client.

Throughout the article, we turn this idea into a novel fast and information-theoretic private information acquisition framework termed, *crowded information acquisition (CIA)*. Rather than individually requesting the desired information to the server, in the CIA, as illustrated in Fig. 1-(b), the clients make a cooperative group, so called a *crowd*, beforehand, and then request the desired information as a group. The effectiveness of the proposed CIA method is shown by examining the overall latency and privacy level under several network environments. Here, the overall latency of a client is defined as the delay that the client experiences from the occurrence of an individual dataset request to the completion

This research was supported in part by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2020-0-01787) supervised by the Institute of Information & Communications Technology Planning & Evaluation (IITP), and in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT). (No. 2021R1A2C2003230).

H. Seo is with Centre for Wireless Communications, University of Oulu, Oulu 90014, Finland (e-mail: hyowoon.seo@oulu.fi). K. Son, and S. Park are with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Korea (e-mail: skrandrew, sangjunpark}@kaist.ac.kr). W. Choi is with the Institute of New Media and Communications and Department of Electrical and Computer Engineering, Seoul National University (SNU), Seoul 08826, Korea (e-mail: wanchoi@snu.ac.kr). *(Corresponding author: Wan Choi)*

of the desired dataset download while guaranteeing the privacy protection. We assume the system is running on the basis of time-division multiple access (TDMA) approach, so that the communication-efficiency can be tested by measuring the latency. The privacy level is an information-theoretic measure that describes how the proposed method is capable of protecting the client privacy. Compared to the conventional approach, referred to as individual information acquisition[1] described in Fig. 1-(a), though it depends on the time overhead for making a crowd, we show that the CIA can retrieve the desired information faster while achieving a higher privacy level in the most of the cases. We also claim that the proposed method is especially communication-efficient compared to the conventional, when there are a large number of clients. The communication-efficiency of the proposed CIA is basically coming from the multicasting gain by reducing redundant information transmitted over the network, compared to the individual information acquisition served with unicasting.

II. System Model

The network under study is composed of a data server and N clients, wherein the clients and server are communicating over wireless. We assume that the clients are located within a circular region of radius R and the distance between the center of the circular region and the data server is $R_s > R$. Consider that the server is storing $K(\geq N)$ independent datasets, i.e., $\mathcal{D}_1, \dots, \mathcal{D}_K$, and clients are making requests for private dataset downloads. Here, the private download describes the case when the dataset download is done without giving a clue on the identity of the dataset to the server which tries to siphon off that identity from the client request.

A. Communication Model

All communications considered throughout the article are half-duplex and done in TDMA. Let $H_{i,j}$ be the channel gain from one $i \in \{s, 1, \dots, N\}$ to another $j \in \{s, 1, \dots, N\}$, $j \neq i$, where s describes the data server and the numbers $1, \dots, N$ describe the client indices. The channel follows the complex normal distribution with zero mean and variance σ^2 ($H_{i,j} \sim \mathcal{CN}(0, P)$), where P is the transmission power. We assume that the channel gains are independently and identically distributed (i.i.d.) and channel reciprocity holds between any two endpoints, that is $H_{i,j} = H_{j,i}$. We define *ϵ-outage transmission time* to describe the successful transmission time in fading channels in a stochastic manner. Note that the data server only transmits datasets of size M_s and clients only transmit request/coordination messages of size $M_c << M_s$, and moreover, transmission power used at the server P_s is relatively larger than that used at the clients $P_c < P_s$ in general. Hence, we define two different ϵ-outage transmission times $\tau_{\epsilon,s}$ and $\tau_{\epsilon,c}$ for the server and clients, respectively.

Assuming a block fading model, where the target transmission time is shorter than the channel coherence time, and thus the channel stays unchanged during the target transmission time, we define $\tau_{\epsilon,s}$ as the minimum τ such that

$$\Pr\left[\frac{M_s}{W_s \log_2\left(1 + \frac{|H_{s,i}|^2}{P_n R_i^\alpha}\right)} \geq \tau\right] \leq \epsilon, \tag{1}$$

for all $i \in \{1, \dots, N\}$, where R_i denotes the distance between the server and the user i, α is the path-loss exponent, P_s is the transmission power of the server, W_s is the bandwidth used by the server and the channel is assumed to experience an additive white Gaussian noise (AWGN) with zero mean and variance P_n. From (1) and the cumulative distribution function (c.d.f.) of the exponential distribution, we have

$$\Pr\left[|H_{s,i}|^2 \leq h_i\right] = 1 - e^{-\frac{h_i}{P_s}} \leq \epsilon, \tag{2}$$

where $h_i = \left(2^{M_s/(W_s\tau)} - 1\right) P_n R_i^\alpha$. Since $R_i \leq R_s + R$, for all $i \in \{1, \dots, N\}$, and from (1) and (2), τ can be bounded as

$$\tau \geq \frac{M_s}{W_s \log_2\left(1 - \frac{P_s}{P_n (R_s + R)^\alpha} \log(1 - \epsilon)\right)}. \tag{3}$$

From the definition, we take the minimum τ as the ϵ-outage target transmission time of each dataset from the server to the client as $\tau_{\epsilon,s} = \frac{M_s}{W_s \log_2\left(1 - \frac{P_s}{P_n (R_s+R)^\alpha} \log(1-\epsilon)\right)}$. Similarly we define the ϵ-outage target transmission time of a client as $\tau_{\epsilon,c} = \frac{M_c}{W_c \log_2\left(1 - \frac{P_c}{P_n (2R)^\alpha} \log(1-\epsilon)\right)}$, where W_c is the bandwidth leveraged for transmission at a client. For the failed transmissions, we suppose that the transmitter compensates for them by retransmissions, where the average transmission time of the server until successful transmission is $\bar{\tau}_s = \tau_{\epsilon,s}/(1 - \epsilon)$ by using the mean of geometric random variables. Similarly, the average transmission time of a client is $\bar{\tau}_c = \tau_{\epsilon,c}/(1 - \epsilon)$.

B. Privacy Model

We focus on the information-theoretic definition of privacy [1] and its variation throughout the article. In the following $I(A; B)$ denotes the mutual information between variables A and B, and $H(A)$ denotes the information-theoretic entropy of A. Supposing $\mathcal{D}_k$ is requested to the server by a client with a query $\mathcal{Q}$ and k drawn uniformly over $\{1, \dots, K\}$, we say that the privacy of the client is perfectly protected when the mutual information is zero such that

$$I(k; \mathcal{Q}) = 0. \tag{4}$$

This perfect privacy, however, is dependent on the cardinality of the set $\{1, \dots, K\}$, which means that the perfect privacy is harder to achieve when K becomes larger. Therefore, in this paper, we define a *privacy level* Π in terms of the conditional entropy that measures the uncertainty of the requested dataset's identity with given query

$$\Pi \triangleq H(k \,|\, \mathcal{Q}). \tag{5}$$

Notice that the perfect privacy is achieved when $\Pi = H(k)$, i.e., $\Pi = \log_2 K$ bits, thereby achieving (4), and identity of requested dataset is disclosed when $\Pi = 0$, which illustrates that there exists no more uncertainty on k.

[1] The individual information acquisition considered as a benchmark in this paper can be seen as the special case of the private information acquisition [1] with a single database server

III. Crowded Information Acquisition

In this section, we propose and investigate *crowded information acquisition (CIA).* The core of the CIA is to aggregate a sufficiently large number of individual dataset requests, and then make a group-wise request for downloading datasets via multicasting. To be specific, the proposed method for the client privacy protection is basically composed of the following three step procedure (see Fig. 1):

(1) Crowding - The crowding is a process of aggregating the dataset requests generated by individual clients. To do this, the clients form a decentralized network and go through process of exchanging the requests. For coordination over a decentralized network, suppose the number of clients N is known to the clients and consider the slotted-ALOHA protocol, which can coordinate the clients to have consensual dataset requests within $\mathcal{O}(N)$ time slots ($\tau_{\epsilon,c}$). Note that the medium access control (MAC) protocols such as slotted-ALOHA can be seen as a solution to the wireless leader (requesting client) election problem. Thus, after N rounds of leader election without replacement, N clients have consensual dataset requests. Furthermore, though the client network is decentralized, it is resilient against faulty nodes since the message exchange is done over wireless and every client is aware of the others' requests [6], [7].

(2) Crowded Requesting - The first client who initiated the crowding[2] sends a query that incorporates consensual datasets which are agreed on downloading by the crowd. In addition, the identities of the members are sent as long as the server only serves registered clients, however, doing this still does not hinder individual request privacy protection.

(3) Crowded Downloading - Third, the server sends the series of requested datasets via multicasting over a shared channel to the members in the crowd.

Therefore, the latency of the CIA incorporates the time required for the three steps, i.e., crowding (*coordination*) T_1, crowded dataset requesting (*uplink*) T_2, and crowded dataset downloading (*downlink*) T_3. Overall, the latency for a client in the crowd to obtain the desired dataset is

$$T_o = T_1 + T_2 + T_3. \tag{6}$$

A. No Privacy Requirement between Members

Now we demonstrate the advantage of the proposed CIA for protecting client privacy. To clearly show the profit of the proposed method, we first consider a scenario where the clients do not have antipathy against disclosing the desired datasets to the other members. Consider N clients, each desiring one of the datasets from the server. Instead of individually requesting the desired datasets, the clients first make a crowd of size N, which is a group of clients that will request the desired datasets together. As aforementioned, the crowding is done by cooperatively exchanging the ind of the desired datasets in advance of requesting and downloading the desired datasets. Based on the slotted-ALOHA approach, the coordination of a crowd of size N is done within $T_1^{(a)} = CN\tau_{\epsilon,c}$, for some constant C. Normally for a vanilla slotted-ALOHA it is known that $C = e$ for sufficiently large N.

After the aggregation, the crowd requests for dataset download. Note if each client uniformly selects $m \leq K$ indices over the index set $\{1, \ldots, K\}$ without replacement, where the selections are independent between the clients, then the number of distinct indices chosen by the clients X_m is distributed according to the shifted binomial distribution with probability mass function (p.m.f.)

$$\Pr[X_m = x] = \binom{K-m}{x-m}(1-p_m)^{x-m} p_m^{K-x}, \tag{7}$$

where $p_m = \left(1 - \frac{m}{K}\right)^{N-1}$, for a positive integer $x \geq m$. Since $m = 1$ in this case, we express the number of the collected distinct dataset indices as X_1, which follows the p.m.f. in (7) with $m = 1$. From the assumption that the size of the query size increases in proportion to the number of requesting datasets, it takes $T_2^{(a)} = X_1\tau_{\epsilon,c}$ for delivering the query to the server.

Upon receiving the request message, the sever multicasts the requested datasets to the clients. Depending on the order of multicasting, the time when a client gets its desired dataset will vary. Thus, we consider the worst case, when the desired dataset comes the latest among all the other requested dataset. Then, it takes $T_3^{(a)} = X_1\tau_{\epsilon,s}$ for the desired dataset download. As a result, we have following proposition.

Proposition 1. *The overall latency the CIA for a client in the crowd of N clients to achieve the privacy level* $\Pi^{(a)} = \log_2 X_1$ *against the server can be expressed as*

$$T_o^{(a)} = (CN + X_1)\tau_{\epsilon,c} + X_1\tau_{\epsilon,s}, \tag{8}$$

for some constant $C > 0$. *Moreover, the average overall latency, with respect to* X_1 *and considering retransmissions, is*

$$\mathsf{E}[T_o^{(a)}] = (CN + g(1))\bar{\tau}_c + g(1)\bar{\tau}_s, \tag{9}$$

while the average privacy level against the server can be approximately expressed by using big-O notation as

$$\mathsf{E}[\Pi^{(a)}] = \log_2(g(1)) - \frac{(g(1)-1)p_1}{2g(1)^2\log_e 2} + \mathcal{O}(K^{-2}), \tag{10}$$

where $g(m) = m + (K-m)(1-p_m)$.

Proof: The equation (9) can be readily derived from the average of binomial distribution, and (10) is from the Taylor expansion of $\log_2 X_1$. ■

From Proposition 1, it can be seen that as the number of members in the crowd N increases, the average achievable privacy level against the server increases. Also note that if all the datasets desired by the members are distinct, the CIA can achieve the privacy level of $\Pi^{(a)} = \log_2 N$ and only a single dataset download is required per each client. Therefore, the proposed method is especially effective in the information acquisition scenarios with a large number of clients in the crowd.

[2] It is reasonable for the initiator to send the group message, since with high probability, the initiator spent less transmission power among the others that went through the multiple rounds of leader election in the crowding phase.

B. Given Privacy Requirement Against Server

In the mean time, consider the case when each client has a privacy level requirement Π^* against the server. Depending on the diversity of the collected indices of the desired datasets during the crowding, the CIA can select one among two predefined strategies. If the number of collected distinct dataset indices X_1 desired by clients, distributed according to (7) with $k = 1$, is sufficiently large so that $\Pi^* \leq \log_2 X_1$, the crowd can follow the same process as done in Sec. III-A. On the other hand, if $\Pi^* > \log_2 X_1$, $2^{\Pi^*} - X_1$ more datasets must be requested in addition to the desired datasets to satisfy the privacy requirement.

Remark 1. *The overall latency of the CIA for a client in the crowd of N clients with given privacy level requirement Π^* against the server is*

$$T_o^{(b)} = \begin{cases} (8), & \Pi^* \leq \log_2 X_1, \\ (CN + 2^{\Pi^*})\tau_{\epsilon,c} + 2^{\Pi^*}\tau_{\epsilon,s}, & \Pi^* > \log_2 X_1. \end{cases} \quad (11)$$

for some constant $C > 0$.

Note that the overall latency $T_o^{(b)}$ averaged over X_1 for given condition $\Pi^* \leq \log_2 X_1$ is the same as (9). Moreover, the probability of having the condition $\Pi^* \leq \log_2 X_1$ can be simply expressed as

$$\Pr[\Pi^* \leq \log_2 X_1] = \sum_{x=2^{\Pi^*}}^{K} \binom{K-1}{x-1} p_1^{K-x}(1-p_1)^{x-1}. \quad (12)$$

C. Given Privacy Requirement Between Members

For a general network, without any given assumptions on relations between the clients, the system must additionally handle the problem that comes from possible privacy leakage during crowding. In such case, the clients must hide the desired dataset by proposing sufficient number of dummy datasets in order to satisfy the given privacy requirement between the members.

Suppose that the privacy level requirement is given by $\Pi^{\#} \leq \Pi^*$ between any pair of members in the crowd as well as the requirement Π^* against the server. Then, each client must choose $2^{\Pi^{\#}} - 1$ dummy datasets uniformly over $K - 1$ datasets that are not desired, and propose $2^{\Pi^{\#}}$ datasets including the desired one when crowding with other members. Moreover, by considering the privacy level requirement against the server, similar to Proposition 1, we have the following.

Remark 2. *The overall latency of the CIA for a client in the crowd of N with satisfying the privacy level requirements Π^* and $\Pi^{\#}$, respectively, is*

$$T_o^{(c)} = \begin{cases} (2^{\Pi^{\#}}CN + X_{2^{\Pi^{\#}}})\tau_{\epsilon,c} + X_{2^{\Pi^{\#}}}\tau_{\epsilon,s}, & \Pi^* \leq \log_2 X_{2^{\Pi^{\#}}}, \\ (2^{\Pi^{\#}}CN + 2^{\Pi^*})\tau_{\epsilon,c} + 2^{\Pi^*}\tau_{\epsilon,s}, & \Pi^* > \log_2 X_{2^{\Pi^{\#}}}. \end{cases} \quad (13)$$

for some constant $C > 0$.

Note that the overall latency $T_o^{(c)}$ averaged over $X_{2^{\Pi^{\#}}}$ for given condition $\Pi^* \leq \log_2 X_{2^{\Pi^{\#}}}$ is the same as

$$\mathsf{E}[T_o^{(c)}] = (2^{\Pi^{\#}}CN + g(2^{\Pi^{\#}}))\bar{\tau}_c + g(2^{\Pi^{\#}})\bar{\tau}_s. \quad (14)$$

Moreover, the probability of having the condition $\Pi^* \leq \log_2 X_{2^{\Pi^{\#}}}$ can be derived by replacing $g(1)$ with $g(2^{\Pi^{\#}})$ in (12).

IV. CIA WITH RANDOM REQUEST ARRIVALS

In the former section, we simply assumed that the clients in the network are requesting datasets at the same instance and making a crowd of size N. However, in the real world, the dataset requests are generated randomly over time, and thus we now consider that the generation of the requests are following the Poisson arrival model at each client. Then the time required for crowding will vary, depending on the request rate.

A. Poisson Request Arrival Model

Consider that the dataset requests are generated following the Poisson arrival process model [8] with the rate λ at each client, which is defined as the average number of requests within a unit time span $\tau_{\epsilon,c}$, and assume that the desired datasets are uniformly distributed over the set $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$ stored at the server. Supposing that the crowd is composed of N dataset requests, the crowding time can be expressed as $T_1^{(d)} = \sum_{i=2}^{N} Y_i \tau_{\epsilon,c}$, where $Y_1, \ldots, Y_N$ are the random variables defined as the inter-arrival times until the N-th request arrival in the network and it is assumed that the coordination time between the clients is assumed to be negligibly small compared to the inter-arrival time.

Note that the inter-arrival time of the Poisson arrival process with parameter λ is known to follow an i.i.d. exponential distribution with parameter λ, i.e., $Y_n \sim \text{Exp}(\lambda)$ for all $n \in \{1, \ldots, N\}$. It can be easily shown that the time span between the first arrival and the N-th arrival is distributed according to the gamma distribution with parameters $N-1$ and λ, i.e., $\sum_{n=2}^{N} Y_n \sim \text{Gamma}(N-1, \lambda)$). Moreover, since the number of distinct datasets requested until N request arrivals is randomly distributed according to the p.m.f. (7) with $m = 1$, we have $T_2^{(d)} = X_1 \tau_{\epsilon,c}$ and $T_3^{(d)} = X_1 \tau_{\epsilon,s}$ for requesting and downloading datasets, respectively.

Proposition 2. *Considering Poisson request arrival with rate λ, the overall latency of the CIA for a client, whose request is in the crowd consisting of N requests and with achieving the privacy level $\Pi^{(d)} = \log_2 X_1$ against the server, can be expressed as*

$$T_o^{(d)} = \left(\sum_{i=2}^{N} Y_i + X_1 \right) \tau_{\epsilon,c} + X_1 \tau_{\epsilon,s}, \quad (15)$$

where X_1 is distributed according to (7) with $k = 1$. Moreover, the average overall latency, with respect to $Y = \{Y_2, \ldots, Y_N\}$, X_1 and considering retransmissions, is

$$\mathsf{E}[T_o^{(d)}] = \left(\frac{N-1}{\lambda} + g(1) \right) \bar{\tau}_c + g(1)\bar{\tau}_s. \quad (16)$$

Proof: Since X_1 different datasets are requested, we have (15) and the average overall latency (16) is from the results in Proposition 1 and the fact that average of $N - 1$ i.i.d. exponential distributed random variables with parameter λ is $(N-1)/\lambda$. ■

B. Poisson Arrivals and Privacy Requirement Against Server

If the target privacy level against the server is given, the clients can stop crowding if there are enough numbers of distinct datasets desired by the members of the crowd. Note that this scenario can be also seen as each dataset request arrives according to the independent Poisson arrival process model with rate λ/K.[3] Suppose that the clients are crowding until 2^{Π^*} distinct datasets are desired in order to satisfy the given requirement. Let $Z_1, \dots, Z_K$ be the random variables denoting the first request arrival time of datasets $\mathcal{D}_1, \dots, \mathcal{D}_K$, respectively, and let $Z_{(1)}, \dots, Z_{(K)}$ be the ascending order statics of $Z_1, \dots, Z_K$, e.g., $Z_{(1)} = \min\{Z_1, \dots, Z_K\}$ and $Z_{(K)} = \max\{Z_1, \dots, Z_K\}$. Since the crowding is done when 2^{Π^*} distinct datasets are desired, the crowding time can be written as $T_1^{(e)} = (Z_{(2^{\Pi^*})} - Z_{(1)})\tau_{\epsilon,c}$. Since the first arrival time of request for $\mathcal{D}_k$ is following an independent exponential distribution, i.e. $Z_k \sim \text{Exp}(\lambda/K)$, the mean of the first order statistic is

$$\mathsf{E}[Z_{(1)}] = \frac{1}{\lambda} \tag{17}$$

and the mean of k-th order statistic, for $k < K$ is

$$\mathsf{E}[Z_{(k)}] = \frac{K}{\lambda}\left(\sum_{j=1}^{k} \frac{1}{K-j+1}\right). \tag{18}$$

Since 2^{Π^*} different datasets are requested to achieve the target privacy level, we can get the overall latency as in the following.

Remark 3. *Considering that requests for each dataset are generated following the Poisson arrival process with rate λ/K, the overall latency of the CIA for a client with achieving the privacy level requirement Π^*, can be expressed as*

$$T_o^{(e)} = 2^{\Pi^*}\tau_{\epsilon,c} + \left(Z_{(2^{\Pi^*})} - Z_{(1)} + 2^{\Pi^*}\right)\tau_{\epsilon,s}, \tag{19}$$

Moreover, the overall latency averaged with respect to $Z = \{Z_{(1)}, Z_{(2^{\Pi^*})}\}$ is

$$\mathsf{E}[T_o^{(e)}] = 2^{\Pi^*}\bar{\tau}_c + \left(\frac{K\sum_{j=1}^{2^{\Pi^*}}\frac{1}{K-j+1} - 1}{\lambda} + 2^{\Pi^*}\right)\bar{\tau}_s. \tag{20}$$

C. Poisson Arrivals and Privacy Requirement b/w Members

For the case when both the against-server and between-client privacy requirements are given, we can use a splitting approach similar to the one used in Sec. IV-B. Owing to the given privacy requirement between the members, we can split the request arrival rate of each dataset to $\frac{\lambda}{K}2^{\Pi^\#}$. Suppose that the clients are crowding until 2^{Π^*} distinct datasets are requested in order to satisfy the given privacy requirement against the server. Let $Z'_1, \dots, Z'_K$ be the random variables respectively denoting the first request arrival time of the datasets $\mathcal{D}_1, \dots, \mathcal{D}_K$ and let $Z'_{(1)}, \dots, Z'_{(K)}$ be the ascending order statics of $Z'_1, \dots, Z'_K$. Then, from Proposition 3, we have the following result.

Remark 4. *Considering each dataset request follows the Poisson arrival process with rate $\frac{\lambda}{K}2^{\Pi^\#}$, the overall latency of the CIA for a client, whose dataset request is in the crowd satisfying the privacy requirements Π^* and $\Pi^\#$, can be expressed as*

$$T_o^{(f)} = 2^{\Pi^*}\tau_{\epsilon,c} + \left(Z'_{(2^{\Pi^*})} - Z'_{(1)} + 2^{\Pi^*}\right)\tau_{\epsilon,s}. \tag{21}$$

Moreover, the overall latency averaged with respect to $Z' = \{Z'_{(1)}, Z'_{(2^{\Pi^*})}\}$ is

$$\mathsf{E}[T_o^{(f)}] = 2^{\Pi^*}\bar{\tau}_c + \left(\frac{K\sum_{j=1}^{2^{\Pi^*}}\frac{1}{K-j+1} - 1}{2^{\Pi^\#}\lambda} + 2^{\Pi^*}\right)\bar{\tau}_{\epsilon,s}. \tag{22}$$

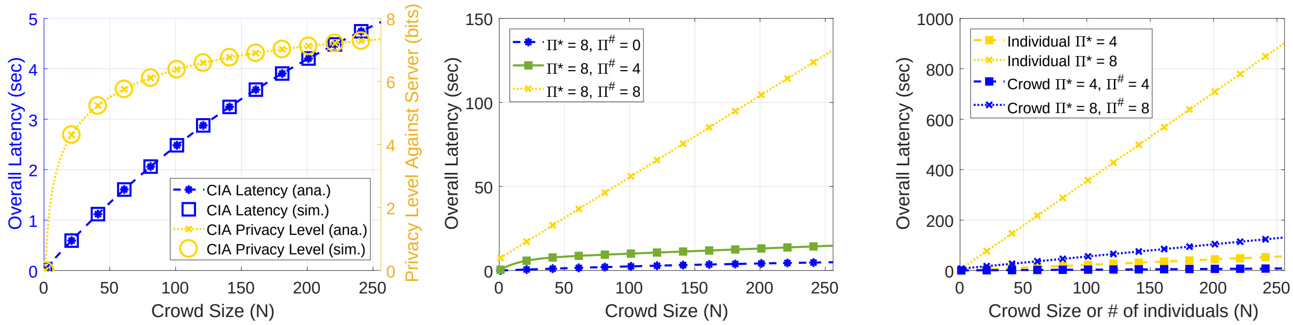
V. Experiments and Discussions

This section verifies the results obtained in Secs. III and IV via numerical experiments. For the experiments, we fix the number of datasets stored at the server to $K = 256$. Moreover, the simulation parameters to reflect the real world communication scenarios are chosen as follows: $R = 500$, $R_s = 1000$ meters, the transmission powers of a client and server are respectively fixed to $P_c = 1$ and $P_s = 100$ Watts, while the thermal noise power is $P_n = 10^{-10}$ Watts. We assume that the request message and dataset sizes are $M_c = 10^3$ bits and $M_s = 10^6$ bits. Note that as the ratio M_s/M_c increases (and correspondingly $\tau_{\epsilon,s}/\tau_{\epsilon,s}$ increases), the efficiency of the proposed CIA, compared to the individual information acquisition, increases. The bandwidth is set $W = 10^8$ Hz and the path-loss exponent is fixed to $\alpha = 3$ considering the urban outdoor scenarios. The outage probability threshold is chosen to be $\epsilon = 0.001$ and the constant is fixed to $C = e \approx 2.718$.

Fig. 2 verifies the results obtained in Sec. III. First of all, Fig. 2a illustrates the overall latency of the CIA with no privacy requirements between the clients, i.e., $T_o^{(a)}$, with respect to the size of the crowd. As shown in Proposition 1, the achieved privacy level against the server increases as the number of members in the crowd increases, while the overall latency is increasing almost linearly as the number of members increases. Note that the overall latency of the conventional individual information acquisition increases quadratically ($\propto N^2$) for the same privacy level against the server as that of the CIA. Fig. 2b illustrates the comparison of the overall latency with perfect privacy requirement against the server ($\Pi^* = 8$) and different privacy requirements between the clients ($\Pi^\# = 0,\ 4\ 8$). Naturally, with larger privacy requirements between the clients, the overall latency increases. However, Fig. 2c shows that even though there are privacy requirements within the crowd, the CIA still outperforms the conventional individual information acquisition. Note that such communication-efficiency comes from the multicasting gain owing to the CIA as mentioned earlier.

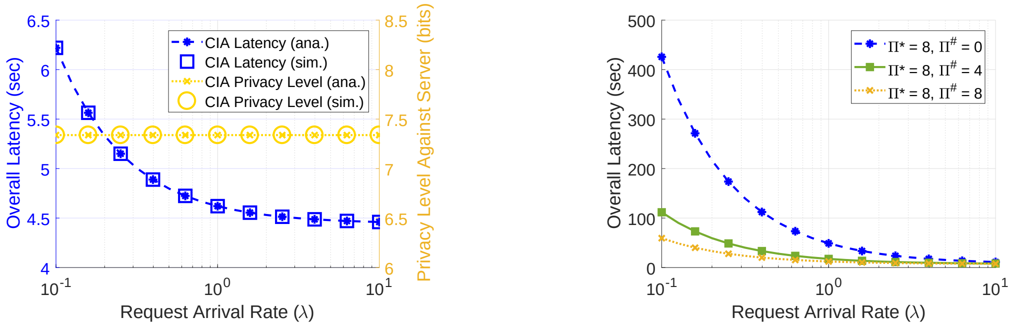
Fig. 3 verifies the results obtained in Sec. IV. First, Fig. 3a illustrates the overall latency of the CIA with Poisson arrivals of the dataset requests $T_o^{(d)}$, with respect to the arrival rate. Furthermore, Fig. 3b illustrates the overall latency of the CIA with perfect privacy requirement against the server and different privacy requirements between the clients ($\Pi^\# =$

[3] Strictly saying, we split (or thin) the original Poisson arrival process in to K independent processes, which is not exactly the same as the original random process.



(a) (b) (c)

Fig. 2. (a) The overall latency and achieved privacy level against the server of the CIA versus the size of the crowd, without privacy requirements between clients, (b) the overall latency with perfect privacy requirement against the server ($\Pi^{*} = 8$) and different privacy requirements between the clients ($\Pi^{\#} = 0,\ 4\ 8$), (c) the overall latency comparison of the individual information acquisition and CIA.



(a) (b)

Fig. 3. (a) The overall latency and achieved privacy level against the server of the CIA versus the dataset request arrival rate λ, (b) the overall latency of the CIA with perfect privacy requirement against the server and different privacy requirements between the clients ($\Pi^{\#} = 0,\ 4,\ 8$) versus λ.

$0,\ 4,\ 8$) versus λ. Interestingly, unlike the result shown in Fig. 2b, for larger $\Pi^{\#}$, the overall latency is smaller for given request arrival rate. This is because if the clients has larger privacy requirements between themselves, making a crowd that achieves the privacy requirement against the server is done faster.

VI. Conclusion

Towards supporting fast and private information acquisitions for a large number of clients, we proposed a novel and communication-efficient privacy protection framework of CIA. We introduced the operational procedure of the CIA, analyzed the CIA under various scenarios, and derived the closed-form expressions of the average overall latency for each scenario. Compared to the conventional individual information acquisition approach, the CIA greatly reduces the overall latency while achieving higher privacy level against the server, which means that the privacy protection is more enhanced against the database server. The proposed CIA is especially efficient when there are a large number of clients in the network and more useful when the network requires user registration so that anonymity-based schemes can not be employed. Our future work will be directed toward a twofold aim considering computational privacy protecting methods and further enhancing communication-efficiency via coded multicasting approach.

References

[1] H. Sun and S. A. Jafar, “The capacity of private information retrieval,” *IEEE Transactions on Information Theory*, vol. 63, pp. 4075–4088, July 2017.

[2] H. Seo and W. Choi, “A stochastic approach in private information retrieval,” in *Proc. IEEE Wireless Communications and Networking Conf. (WCNC)*, pp. 1–6, Apr. 2018.

[3] K. Banawan and S. Ulukus, “The capacity of private information retrieval from coded databases,” *IEEE Transactions on Information Theory*, vol. 64, pp. 1945–1956, Mar. 2018.

[4] K. Banawan and S. Ulukus, “The capacity of private information retrieval from byzantine and colluding databases,” *IEEE Transactions on Information Theory*, vol. 65, pp. 1206–1219, Feb. 2019.

[5] H. Sun and S. A. Jafar, “The capacity of symmetric private information retrieval,” *IEEE Transactions on Information Theory*, vol. 65, pp. 322–329, Jan. 2019.

[6] H. Seo, J. Park, M. Bennis, and W. Choi, “Consensus-before-talk: Distributed dynamic spectrum access via distributed spectrum ledger technology,” in *Proc. IEEE Int. Symp. Dynamic Spectrum Access Networks (DySPAN)*, pp. 1–7, Oct. 2018.

[7] H. Seo, J. Park, M. Bennis, and W. Choi, “Communication and consensus co-design for distributed, low-latency and reliable wireless systems,” *IEEE Internet of Things Journal*, vol. 8, pp. 129–143, Jan. 2021.

[8] R. G. Gallager, *Stochastic Processes: Theory for Applications*. Cambridge University Press, 2013.