Cost Analysis of Query-Anonymity on the Internet of Things

by

Abdul Kadhim Hayawi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2017

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Dr. Changcheng Huang Professor
Supervisor(s)	Dr. Pin-Han Ho Professor
Internal Member	Dr. Kshirsagar Naik Professor
Internal-external Member	Dr. Bernard Wong Associate Professor
Other Member(s)	Dr. Lin Tan Associate Professor

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

A necessary function of the Internet of Things (IoT) is to sense the real-world from the fabric of everyday environments. Wireless Sensor Networks (WSNs) are widely deployed as part of IoT for environmental sensing, industrial monitoring, health care, and military purposes. Traditional WSNs are limited in terms of their management and usage model. As an alternative paradigm for WSN management, the sensor-cloud virtualizes physical sensors. While this model has many benefits, there are privacy issues that are not yet addressed. The query-anonymity arises when the client wants the destination physical sensor-node to be indistinguishable from other potential destinations. In particular, we consider the k-anonymous query scheme in which the query destination is indistinguishable from other $k-1$ probable destinations, where k is the offered level-of-anonymity. Moreover, we are interested in a communication-based approach in which the client is required to send k queries to at least k destinations including the node of interest in order to achieve a level-of-anonymity k. Thus, the communication-cost increases with the level-of-anonymity k. Consequently, there is a natural trade-off between the offered query-anonymity and the incurred communication-cost. The analysis of such trade-off is the main problem we address in this work.

We firstly aim at a novel theoretical framework that quantifies the security of a general k-anonymous query scheme. Towards that, we adopt two approaches based on well-known security models namely, ciphertext indistinguishability under chosen plaintext attack (IND-CPA), and information theoretic notion of perfect secrecy. Next, we provide a construction of a secure k-anonymous query scheme, and introduce its detailed design and implementation, including the partition algorithm, anonymity-sets construction methods, query routing algorithm, and querying protocol. Then we establish a set of average-case and worst-case bounds on the cost-anonymity trade-off. We are committed to answer two important questions: what is the communication-cost, on average and in the worst-case, that is necessary? and what is the communication-cost that is sufficient to achieve the required secure query k-anonymity?

Finally, we conduct extensive simulations to analyze various performance-anonymity trade-offs to study the average and worst-case bounds on them, and investigate several variations of anonymity-sets constructions methods. Confirming our theoretical analysis, our evaluation results show that the bounds of the average and worst-case cost change from quadratic asymptotic dependence on the network diameter to the same dependence on the level-of-anonymity when the later surpasses the former. Furthermore, most of the obtained bounds on various performance anonymity trade-offs can be expressed precisely in terms of the offered level-of-anonymity and network diameter.

Acknowledgements

First of all, I owe a special debt of gratitude to my thesis supervisor Prof. Pin-Han Ho, for his kind and expert guidance, advise, help, and support that was pivotal for my research work. I would like to thank Prof. Pin-Han Ho who relentlessly inspired me to pursue high quality research, believed in me, refined my efforts, and critically reviewed my work. He always taught me to aim high and not to give up. I am deeply grateful to him.

I would like to thank all professors, friends, classmates, and research group members at University of Waterloo for their help, support, and valuable discussions. I thank Prof. Pin-Han Ho, Prof. Doug Stinson, Prof. Otman Basir, and Prof. Weihua Zhuang for their superb teaching during my PhD studies.

I am grateful for my thesis examining committee Prof. Pin-Han Ho, Prof. Kshirsagar Naik, Prof. Changcheng Huang, Prof. Lin Tan, and Prof. Bernard Wong for kindly offering to read my thesis, and to give their insightful comments. It is honor to have you all.

At last, I would like to express my deep respect and love to my parents, my wife, my daughter, my sons, my brothers, my sisters, and their families for their constant support, and sacrifices.

Dedication

To my dearly beloved family

Table of Contents

List of Tables

List of Figures

List of Notations

k : level-of-anonymity
n : wsn network size
d : network diameter
$G(V, E)$: a graph with vertices set V and edge set E
v : vertex
DAS : Disjoint Anonymity-Sets scheme
$IND - CPA$: ciphertext indistinguishability under chosen plaintext attack
r : radius of equivalent circular cloaking area
d_{max} : maximum cloaking distance
$H(x)$: Shannon entropy
H_{max} : maximum Shannon entropy
$\mathbb{Q}$: the set of all possible queries
q : a single query
q_i : ith query
Q : discrete random variable denoting the query value
v_d : true destination node of a query
π : the partition algorithm
T : the anonymity transformation
T^{-1} : the anonymity inverse transformation
s : anonymity-set
$\mathbb{S}$: the collection of all possible anonymity-sets
$Pr[.]$: probability of
$|s|$: size of anonymity-set
Q_T : the set of all possible queries at the input of the anonymity transformation
$\mathbb{Q}_{\mathbb{T}}$: the collection of Q_T
Q_k : the subset of indistinguishable queries
S_b : the set of bogus destination queries
$\mathcal{A}$: eavesdropping adversary

$\mathcal{C}$: the challenger or client
$\mathcal{S}$: the sensor-cloud
$G(A, T)$: the adversarial indistinguishability game with adversary $\mathcal{A}$ and anonymity transformation T
$s \setminus Q_T$: the set difference between s and Q_T
$\Theta(.)$: Big Theta notation of tight bound
$O(.)$: Big O notation of upper bound which may or may not be tight
$\Omega(.)$: Big Omega notation of lower bound which may or may not be tight
$o(.)$: little o notation of not tight upper bound which is the complement of $\Omega(.)$
(i, j) : ordered pair Cartesian coordinates of a square grid WSN node
$v_{1,1}$: the root node of WSN which is the entry point to communicate with WSN
d_m : Manhattan distance
$ascm$: anonymity-sets construction method
FSS : First Set Spread anonymity-sets construction method
ES : Equal Spread anonymity-sets construction method
RS : Random Spread anonymity-sets construction method
v_1 : the first node of the anonymity-set from the source with coordinates (x_1, y_1)
$v_{|s_j|}$: the last node of the anonymity-set from the source with coordinates $(x_{|s_j|}, y_{|s_j|})$
t : comb-like spanning tree
P : Hamiltonian path
BFS : Breadth First Search algorithm
h_j : source-route header information
ID_v : WSN node ID value
$Action_v$: WSN node action value
R : a list that stores piggybacked response data
ρ : a query packet
V_d : discrete random variable whose value equal to the destination node v_d
c_a : average-case communication-cost
c_w : worst-case communication-cost
ROI : Return-On-Investment
d_c : cloaking distance
c_0, c_1, c_2 : asymptotic constants of $\Theta(.), \Omega(.), O(.)$ respectively

Chapter 1

Introduction

The Internet of Things (IoT) introduces a paradigm shift in networking that aims to connect real-world things to the Internet. These are things that collect information with potential value for information consumers. These IoT candidates are embedded with computing capabilities, communication protocols, and sensors to blend into everyday environments and seamlessly bridge the gap between people, places, and things over the Internet [1].

Inferring real-world events with sensing devices is an integral part of the Internet of Things (IoT), where virtualization of the real-world things via cloud services is a norm due to their huge volumes and heterogeneity. Cloud services are widely used to virtualize the management and actuation of the real-world on the Internet of Things (IoT). Though virtualization simplifies the control and management of IoT, it raises serious privacy concerns when a client issues queries to a virtual entity that is nonetheless handled by an untrusted cloud service provider. Consequently, query-anonymity has become a critical issue to all the stakeholders which are related to assessment of the dependability and security of the IoT system.

1.1 Sensor-Cloud-Based IoT Systems

Although many real-world things have valuable information to share, it is a challenge in equipping them with all the required capabilities to connect to the Internet. The wide adoption and enhancement of technologies like RFID, cloud computing, and Wireless Sensor Network (WSN) have enabled the virtualization of the real-world things [1, 2]. For instance, things with embedded sensors provide sensory data that are accessed as Web services in the cloud and create the possibility of composing different services to infer important knowledge about physical environments [2]. The sensed knowledge is used to enhance the capability of the things that do not have sensing capability but are in the same location as those things that have sensing capabilities.

Things in the physical world increasingly become fully qualified members of the Internet. These IoT candidates publish themselves and query each other for information and the interactions are available for the whole world to see. Today, most IoT candidates are either pure sensing devices or with sensing devices embedded within [1]. Wireless Sensor Networks (WSNs) are widely deployed as part of IoT for environmental sensing, industrial monitoring, health care, and military purposes. Traditional WSNs are limited in terms of their management and usage model [3]. As an alternative paradigm for WSN management, the sensor-cloud virtualizes physical sensors. The sensor-cloud provides sensory data to a plethora of potential smart applications via the provision of sensing-as-a-service [2, 4]. The control and management of thousands of these things via cloud services have been a norm, and the virtualization has further enhanced their capabilities and semantics [5, 6].

1.2 Cost-Anonymity Trade-Off Problem

While IoT ubiquitously provides flexible, scalable, and real-time communications with the physical world, there are increasing security and privacy concerns due to such ubiquity. The exposure of real-world environments in the virtual world requires the IoT candidates to have the option of being anonymous. A common situation is when a client queries a physical sensor, the corresponding virtual sensor is queried first, and the query is passed to the cloud as shown in Fig. 1.1, where the owners of the physical WSN and the cloud are decoupled from the client (by the virtual sensor). This significantly impairs the trust factor of the clients.

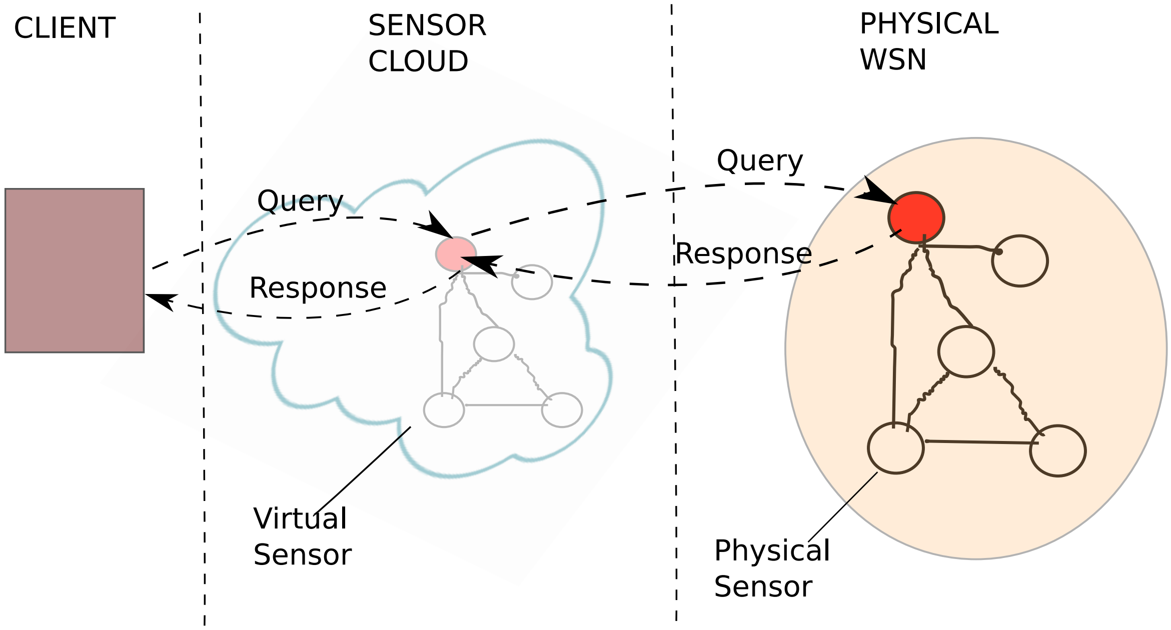


Figure 1.1: The scenario we consider in which a client wants the destination physical sensor to be anonymous in the eye of a passive eavesdropper.

The above problem can be mitigated by launching a k-anonymous query scheme [52, 53, 7, 8, 9], which attempts to make the query destination indistinguishable from other $k-1$ probable destinations, where k is the offered level-of-anonymity. An illustration of k-anonymity is shown in Fig. 1.2. Although extensively investigated in the past, the quantification of security of a general k-anonymous query scheme in large-scale WSN is missing; and the trade-off between the gained security/privacy and incurred communication overhead is not well investigated.

Motivated by the above observations, this work firstly aims at a novel theoretical framework that quantifies the security of a general k-anonymous query scheme. We consider on-demand queries where a client decides when to query the WSN to acquire sensor readings, instead of event-driven or time-driven queries [8]. Secondly, since a k-anonymous query sends additional $k-1$ queries in order to achieve a level-of-anonymity of k, the incurred communication-cost increases with the level-of-anonymity. This trade-off has been recognized in the literature as essential between the offered query-anonymity and the incurred communication-cost [8]. In this work, we go one step further by establishing a set of average-case and worst-case bounds on such trade-off. We are committed to answer two important questions: what is the communication-cost, on average and in the worst-case, that is necessary, and what is the communication-cost that is sufficient to achieve the required query-anonymity? Targeting large scale network of things (IoT) that form WSNs, our cost analysis is asymptotic, i.e. it scrutinizes the lower and upper bounds on the average and worst-case communication-cost to achieve a specific level of secure query-anonymity.

The system models employed in the formulation of our problem, including the trust model, the WSN network model, and the threat model are presented in the following sections.

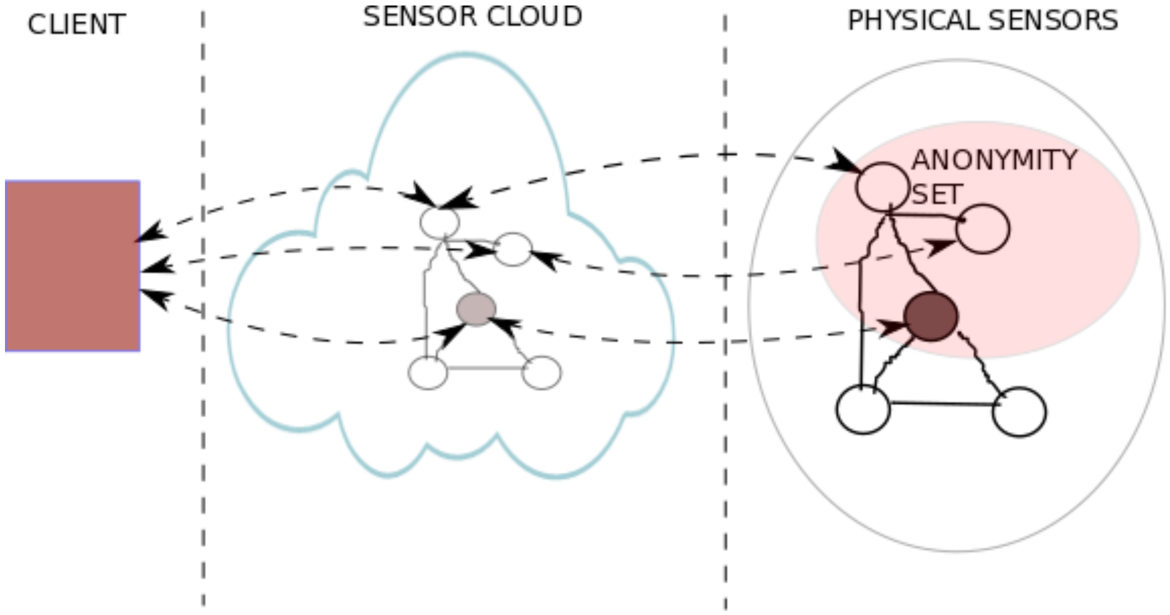


Figure 1.2: An example of k-anonymous query scheme. The client queries three nodes including its true destination (shaded) in order to achieve a level-of-anonymity $k = 3$.

1.2.1 Trust Model and Application Scenario

We consider a Large WSN conglomerate owned by multiple operators, and the WSN could be in any type like terrestrial, underground, and underwater networks for gathering environmental data. The clients query the physical sensors for the sensed data by accessing services provided by the sensor-cloud providers which could be governments, corporates, or academia. The various stakeholder entities in the scenario include the operators who own the WSNs, the cloud service providers, and the clients that query the sensors. Their relation is that the cloud service providers lease resources from the WSN operator/owner to offer the sensing-as-a-service to their clients [2]. Thus, it is the cloud service providers who face their clients but are with limited knowledge on the sensor-cloud operations. Many applications of this model are foreseen in areas like smart cities, smart transportation, and smart manufacturing. The clients use the sensor data for different purposes like traffic management, resource exploration, environmental protection, air pollution control, and civil infrastructure monitoring, etc.

With the above mentioned scenario, we consider the *trust-none model* where a client trusts no other entity. However, the clients are concerned about the privacy of their interest, queries, and data access patterns which may be compromised because of untrusted sensor-cloud owners or other competing clients. One of the appealing characteristics of trust-none model is that all the secrecy transformations are performed by one party namely, the owner of the sensitive information (the client). The reason for that is the owner of sensitive information shares none of her sensitive information with other components of the system; therefore she is the only one who can operates on these information. The trust-none model is so powerful that it allows the owner of the sensitive information to achieve her purpose without revealing any sensitive information to other components of the system.

The k-anonymous query scheme furnishes an interesting application of trust-none model in which the client (owner of sensitive information) need not to trust any other party of the system. In our settings, the sensitive information is the ID number of the distention node that is being queried. Consequently, cryptographic mechanisms which is based on secret key sharing is of no benefit here since there is no trusted entity in the system which the client can share secret with. On the other hand, to allow the destination nodes to respond faithfully to the k-anonymous query, they have to be able to understand it. Hence messages (queries and their responses) must be sent anonymously in plain text which makes the case for communication-based anonymity schemes.

An interesting application scenario is the deep-sea oil exploration, where an oil company (client) is interested in querying the deep sea WSN that is owned by operators who are different from the cloud service provider [8]. Certainly the oil company (client) is not willing to share its interest with other stakeholder entities. It follows that a trust-none model is such that the client treats the WSN, the cloud service provider, and other clients as untrustworthy adversaries.

1.2.2 Network Model

Since query-anonymity solutions in a large-scale WSN are considered, we group the WSN sensor nodes into clusters whose head are elected or pre-assigned by network designer [10]. On the contrary to the flat topology, the clustering-based hierarchical topology performs well in large-scale WSN [10]. This is mainly because of the reduced energy, computation, and communication requirements of sensor nodes in each cluster. The clustering protocols in WSN are energy-efficient in order to prolong the battery life of sensor nodes and the whole WSN lifetime. In addition, they enhance the network coverage, operation and management. For example, the cluster-heads schedules activities, such as transmission and reception time, in their clusters so that cluster members switches to low-power sleep mode more often. Clustering helps to localize the intra-cluster routing messages exchanges within each cluster, and limits the inter-cluster communications to cluster-heads lowering transmission retries and redundant exchanges. Furthermore, cluster-heads can aggregate data collected from sensor nodes in their clusters to further conserve the communication bandwidth and energy consumption. Data aggregation involves combing data from different sensor nodes in the cluster using functions such as min, max, average, and suppression of redundant data. Data aggregation cuts down on energy since computation is an energy-saver as compared to radio data transmission. All of the proceeding advantages of clustering result in reducing the rate of energy consumption for each sensor node in WSN, and make clustering the topology of choice when scalability is the main network design objective [10, 11]. Examples of clustering routing and election protocols designed for WSN are: Low-Energy Adaptive Clustering Hierarchy (LEACH) [12], and Hybrid Energy-Efficient Distributed clustering (HEED) [13].

As in prior work on WSN query anonymity [8, 14], our graph model abstracts the intra-cluster routing among sensor nodes inside each cluster since these details are not relevant to our work. However, the cluster-heads form a second tier network upon which we build a routing tree to facilitate inter-cluster communication. As we will show in Chapter 5, the inter-cluster routing affects the semantics of the incurred communication-cost to achieve a specific query-anonymity level.

Accordingly, the WSN is modeled as undirected connected graph $G(V, E)$, where V (vertices) are a number of n cluster heads joined by links or edges E. For every two nodes $v_i, v_j \in V$, it holds that the edge $(v_i, v_j) \in E$ if and only if whatever v_i transmits, v_j can always receive, i.e., v_i is adjacent to v_j. It is important not to confuse WSN sensor nodes with the nodes in the graph, where the nodes in the graph are cluster-heads that are richer in resources compared to their cluster members and serve as gateway for relaying/broadcasting the clients' queries. We consider all cluster heads to be equal in power, computation, and communication capabilities.

1.2.3 Attacker Model

The adversary in our setting is an unconditional global passive eavesdropper who is able to monitor and analyze all the message exchange between the client and the WSN but does not actively alter them. This is an often adopted adversarial model called the *semi-honest* or *honest-but-curious* adversary [15]. The adversary is global and strong enough to have access to the entire WSN. When there is no limitation on the computational power of the adversary or on her ability to collude with other component of the system, it is unconditional.

We follow Kerchhoffs's principle or Shannon maxim of security by transparency [16]: the adversary knows the whole anonymity scheme including the details of the anonymity transformations. However the adversary is able to conduct cipher-text only traffic analysis attack. That is the attacker has access only to the output of anonymity transformation (the anonymity-set) when it is executed.

1.3 Thesis Motivations and Contributions

The first aim of this work is to review and categorize the major trends in the field of anonymous communication (see Chapter 2). Our contributions in this area are two folds:

1. We propose a cost-based classification paradigm that groups the most influential anonymous communication schemes into two broad classes: crypto-based and communication based systems. Our classification of the anonymous communication systems into crypto-based and communication-based shed light on the fact that little of research work has been done on the communication-based system than the crypto-based systems. Furthermore, we show that the trade-off between latency and anonymity is the main issue in the first category, whereas the main problem in the second is the trade-off between incurred communication-cost and offered anonymity.

2. We also discuss how anonymity metrics and notions have been developed to evaluate the offered security properties of anonymous communication schemes.

The second aim of this work is to propose a theoretical framework that quantifies the security of a general k-anonymous query scheme (see Chapter 3) [14, 17, 18]. The trade-off between communication-cost and anonymity in communication-based anonymous schemes is an area that requires more attention. Since the cost (computational, delay or communication overhead) are usually measurable, measuring anonymity and having the right metrics and notions are the more challenging area of research. Indeed the absence of such a theoretical framework that captures the practical concerns of k-anonymous query scheme motivated this work. The secure query k-anonymity notion proposed in this work is proved to be helpful in the design, analysis of trade-off between query-anonymity and communication-cost, and evaluation of provably secure query-anonymity schemes. We summarize our contributions in this area as follow:

1. We propose an unconditional notions of secure query k-anonymity on the basis of two well-known security models namely, ciphertext indistinguishability under chosen plaintext attack (IND-CPA), and information theoretic notion of perfect secrecy. Our notions formally define what it means to be *secure* for k-anonymous query schemes. They also scrutinize the practical design aspects of secure k-anonymous query scheme.

2. We unify the indistinguishability and perfect secrecy characterization of anonymity with the standard and intuitive concept of anonymity-set in a natural way.

3. We show that constructing equi-probable anonymity-sets is not only sufficient for achieving a certain level-of-anonymity, but also the cheapest manner in which to achieve it.

The third aim of our work is to analyze the trade-off between incurred communication-cost and offered secure query k-anonymity (see Chapters 4, 5, and 6) [17, 18, 19]. Although this trade-off has been recognized in the literature as essential [8], it is not well-investigated. Specifically, our contributions in this area are summarized in the following:

1. We provide a construction of a secure k-anonymous query scheme that is the Disjoint Anonymity-sets (DAS) scheme. DAS satisfies our definition of secure query k-anonymity, and it is agnostic to application and topology.

2. We propose a specification paradigm for the design and implementation of DAS for privacy preservation in the presence of unconditional eavesdropping adversary including its partition algorithm, anonymity-sets construction methods, query routing algorithm, and querying protocol.

3. We establish a set of average and worst-case asymptotic bounds on cost-anonymity trade-offs. We show that the network diameter plays the role of a system-wide inflection point ($k = \Theta(d)$) at which the asymptotic growth of the average and worst-case communication-cost changes from d^2 into k^2 dominance. Additionally, the return-on-investment ROI is at its highest values at this point.

4. We introduce performance metrics to measure the offered location-anonymity namely, the radius of cloaking area r, and the maximum cloaking distances d_{max}.

5. We show that most of the bounds on various performance-anonymity trade-offs are functions of the level-of-anonymity k, and network diameter d.

1.4 Organization of The Thesis

The rest of the thesis is organized as follows. Chapter 2 presents a literature survey on the anonymous communication systems comparing the two prominent approaches namely, the crypto-based, and the communication-based that is our field of study. We also reviews different approached to define and measure anonymity. In Chapter 3, we introduce two novel theoretical frameworks to formally define the secrecy of a general k-anonymous query anonymity scheme based on ciphertext indistinguishability, and information theoretic notion of perfect secrecy models. Chapter 4 firstly presents an application and topology agnostic construction that satisfies our notion of secure query k-anonymity presented in Chapter 3, then we provide the design and implementation details of it. In chapter 5, we analyze the trade-off between the communication-cost and query-anonymity in a sensor-cloud-based IoT setting that utilizes square grid WSN. Our cost analysis shows that the behavior of the secure k-anonymous query scheme is shaped by two variables, namely the level-of-anonymity k, and the network diameter d. Simulation results are given in chapter 6 that confirm our theoretical assertions. We finally conclude the thesis in Chapter 7, and discuss potential future work.

Chapter 2

Cost-Based Classification of Anonymous Communication Schemes, and Their Anonymity Notions - a Survey

In this survey, we review the most influential schemes of anonymous communication, and compare them to our own. We are focusing on the design concepts and major weaknesses that lead to attacking these systems by passive and active attackers.

Considering anonymous communication schemes which deal with sending and receiving of messages anonymously, anonymity is usually used to hide the identities of the actual sender, receiver or both within anonymity-set of other possible senders and/or receivers. They are called *sender anonymity*, *receiver anonymity*, and *relationship anonymity* respectively.

The first anonymous communication scheme is the Mix-Net proposed by David Chaum in 1981 [20]. The Mix-Net hides the relationship between senders' identities on their input, and receivers' identities on their output. Chaum also introduced the notion of anonymity-set in the DC-Net in 1988 [21]. The set of all possible subjects that the anonymous subject belongs to. For example, for the case of sender/receiver anonymity, it is the set of all possible senders/receivers of the anonymous message.

Since the introduction of Mix-Net and DC-Net, different anonymous communication schemes and protocols have been proposed to provide sender, receiver, or relationship anonymity. However, based on the cost-based classification we propose in this chapter, they fall mostly into two broad classes of anonymous communication schemes namely, crypto-based which follows Mix-Net approach, and communication-based which is built on DC-Net idea. Anonymous communication has become an extremely active area of research with wide range of applications. Additionally, a variety of countermeasures have been investigated to protect against possible attacks that compromise the anonymity solutions.

As we mentioned in section 1.3, the contributions of this chapter are as follows. We proposed a classification paradigm for the most influential anonymous communication schemes based on the main component of the cost incurred by the anonymity solution. Thus we classify them into *crpto-based schemes* when the cost is mainly the computational-cost of the cryptographic technique used, and into *communication-based schemes* when the communication overhead cost constitutes the dominant part of the anonymity solution cost. Our own scheme falls under the communication-based class. The chapter also presents an overview of the anonymity notions and metrics that are used to evaluate these schemes.

2.1 Crypto-Based Anonymous Communication Schemes

In this section, we describe the most prominent cypto-based anonymous communication schemes. As we will show, the main issue in this class of anonymous communication scheme is the trade-off between latency and offered anonymity.

2.1.1 Mix-Net Scheme

The Mix-Net consists of intermediate nodes acting as proxy servers that utilize public key cryptography to hide the relationship between senders' identities on their input and receivers' identities on their output. The sender encrypts its message using the public key of the receiver, and then encrypts the result with the public key of the proxy server that forwards the inner encrypted message to the receiver after mixing it with other messages.

The local attacker who observes the messages at the input and output of the mix server is unable to link them, hence the Mix-Net provides unlinkability. However the single mix network introduces a single point of failure, and single point of trust in the system. A feature that is undesirable in most of the high availability systems. To ensure reliability, multiple mix servers are cascaded to route the messages between the sender and receiver. The sender creates multiple layers of encryption starting with encrypting her message using the public key of the receiver. Then she encrypts the resulting ciphertext using the public key of the last mix server, which is the closest to the receiver, and continues backward to the first mix server along the message path. On top of its legacy as a keystone, the high-latency cascade of these proxy servers is essentially the principle of onion routing [22], which is implemented in the low-latency Tor scheme [23].

Two main approaches are proposed in the literature to choose the message path namely:

free routes which allows the sender to define her own message path, as the one used in our source routing algorithm, and *mix cascades* in which the message path is chosen from a predefined set of routes [24]. The *free route* provides better failure tolerance in the event of message path failure since it gives the user more freedom in choosing the best route [25].

The *flushing algorithm* or *batching strategy* [26, 27] is used to decide how, and when to forward the messages out of a mix server. It is a fruitful area of research that is resulted in several designs that falls into two broad categories namely, deterministic and probabilistic. Examples of the deterministic algorithms are: the Threshold, Timed, and Threshold-and-Timed mixes, and examples of the probabilistic algorithms are: Stop-and-Go, and Binomial mixes. The Threshold strategy defines a threshold n on the number of mixed messages after which the mix server starts forwarding message to the next server or to the receiver [20]. In the Timed mix, the messages are outputted out of the mix after a specified time limit t [26], while the *Threshold-and-Timed* mix combine both of the above strategies together [26]. The *Stop-and-Go* Mix forwards the message after an exponentially distributed time delay [28]. The *Binomial* Mix selects the forwarded messages according to binomial distribution [27].

Several attacks against Mix-Net have been devised in the literature. The *flooding* attack or $(n-1)$ attack is an active attack against Threshold mix. In this attack, the adversary delays one message from entering the mix, and floods it with bogus messages till it flushes. Then the adversary allows the delayed message to enter the mix within a flood of dummy messages to recognize it after encryption [26]. The *trickle* active attack uses similar concept of *flood* attack but it targets Timed mixes and uses no dummy messages [26]. The *flooding* and *trickle* use similar attack strategy which is *"blending"* attacks [29]. While the *blending* attacks succeeded against deterministic flushing algorithms, it is more difficult to compromise the probabilistic flushing algorithms. Since original mix design is based on public cryptography using RSA algorithm [30], researchers described active attacks that make use of the fact that RSA is not a randomized encryption. Consequently, similar plaintexts at the input of the mix will result in similar ciphertexts at the output. In addition, a plaintext with known ciphertext will not be anonymous at the output of the mix [31]. As a countermeasure, the input and output messages are made to be equal size using random padding, and ElGamal public key cryptosystem is used instead of RSA since it provides semantic security [32].

The anonymity scheme studied in this thesis is different from Mix-Net in several angles. Unlike our trust-none model, which uses no third party component, Mix-net is a proxy-based system in which a mix server acts as a performance bottleneck, and a single point of failure when an adversary is in control of the proxy, as we show in the proceeding paragraph. Mix-net is vulnerable to the collaborations among mix serves on the contrary of our scheme whose offered anonymity is not affected by the corruption of the system components other than the client herself. Mix-Net relies on public key cryptography [33] to perform its main function which cannot be readily applied to the WSN scarce resources environment. Mix-Net has an advantage in terms of local adversary who can see only the previous and next mix server, while our scheme offers almost similar level-of-anonymity against both of the local and global adversary. Compared to Mix-Net, our scheme incurs much less network latency and computational-cost, but higher communication-cost.

2.1.2 Tor Scheme

Tor is the second generation of onion routing [23]. Onion routing uses cascades of mix servers to provide anonymity for TCP-based communication [22]. The onion consists of layers of encryption that covers the source's address and the message. Every Onion router

(OR), on the path between client and destination, peels off a layer, and appends a random padding to maintain a constant size onion before forwarding to the next OR. The constant size onion prevents leaking information about the position of the OR on the client path to destination due to onion size change.

Tor is optimized for anonymous Web browsing and real time applications to be a low latency anonymous communication scheme. This does not come for free since it relies on forwarding the message in real time without doing a proper mixing.

Tor is a circuit-based communication scheme that utilizes hybrid encryption, symmetric and public key cryptography. Tor circuit establishment starts when the client download a list of onion routers (ORs) from a directory server, typically three ORs. Then the client creates tunneled secure channels with the three ORs using the Transport Secure Layer protocol (TLS) [34]. Afterward the client negotiate three secret keys with the ORs using authenticated Diffie-Hellman key agreement protocol. At this point the client can communicate anonymously over the Internet in real time because each OR knows only its direct adjacent ORs in Tor circuit.

Tor defends against attack that tried to insert malicious ORs in the client path by using dedicated directory servers. This approach is similar to ours that uses source routing instead. However an adversary can compromise the security of Tor by controlling or just watching the first and the last ORs. This is true because Tor, which is designed for Web users, trade anonymity for Web usability and cheap of operation for Internet users [23]. That is unlike Mix-Net, Tor maintains low latency anonymous communication by passing messages in real time and not providing a heavy traffic cover in the mixing process. This vulnerability opens the door to traffic analysis family of attacks in which the adversary intercepts multiple outputs of anonymity scheme in order to exploit redundant patterns in long-lived transactions using statistical methods, and subvert the provided anonymity.

Examples of these attacks are: *intersection*, *predecessor* or *statistical disclosure*, and *traffic correlation*. In *intersection* attack [35, 36, 37, 38], the adversary intersects multiple outputs of the anonymity scheme in order to identify the anonymous subject. In *predecessor* or *statistical disclosure* attack, the adversary counts how often a specific destination has been accessed by the suspected anonymous source for sender anonymity, and vice versa for receiver anonymity, over multiple runs of the anonymity scheme [39, 40]. Such traffic analysis attacks exploit the fact that the scheme leaks information to the adversary who is able to provide not-equally likely probability values to the possible subjects (senders/receivers), and reveal the victim's identity. *Statistical* attack can also be viewed as an intersection of multiple outputs of anonymity scheme to compromise the anonymity of the most frequent common subject among them. Thus, they are used interchangeably in the literature, and in this work. The defence against *intersection* and *statistical* attack is an open problem that is difficult to solve efficiently [41, 42]. Specific to Tor, the adversary mounts *intersection* attack to reveal victim's identity by monitoring how the sets of active Web users, which provides a cover to the victim's traffic, decreases in size by intersecting them over multiple runs of the scheme. Whereas in *statistical* attack, she simply counts how often a Web resource has been accessed by the same Tor user, and identifies the most frequent user as the initiator. During *traffic correlation* attack [43], the adversary correlates timing and other traffic flow characteristics to undermine the anonymity scheme.

An instance of these attacks is demonstrated for the case of using Tor as a mechanism for location-hidden services via rendezvous points in [44]. Countermeasures to combat this attack is based on requiring the hidden server to always connect through "entry gaurds" [44], or the client to first connect to a trusted "valet services node" [45]. Therefore The selection of the first and last ORs is very crucial for the security of Tor.

In contrast to Tor, our scheme requires no trusted nodes, provably secure against global eavesdropping adversary, and incurs much less network latency and computational-cost, but higher communication-cost. Our scheme precludes all types of traffic analysis attacks including intersection and statistical disclosure attacks by having disjoint outputs (anonymity-sets) of the anonymity scheme so that all subjects (members of anonymity-sets) are equiprobable.

2.1.3 Remailer Schemes

Remailer schemes aim at providing sender and receiver anonymity for email users. Being Mix-Net-based, they are high latency store, mix then forward systems which suits non-interactive email application. Three Remailer generations have been developed by researchers along the years, namely: Type I, Type II, and Type III remailers.

Type I is the Cypherpunk [33] which consists of cascade of systems that strip the identities of emails before forwarding them to their final destination. The Pretty Good Privacy (PGP) public key cryptography protocol is used for encryption to defend against eavesdropper. The users' actual identities are not kept in a database which is an important improvement compared to its ancestor anon.penet.fi remailer.

Since PGP does not hide the message size, the message size leaks information to the adversary about the position of the remailer. The reusable reply blocks, which is used for anonymous email reply, is another weak point that an active attacker can use to trace the recipient path. Type II remailer which is the Mixmaster system [46] fixed the problem of the variable message size of Type I, but it kept the reusable reply block vulnerability.

Type III is the Mixminion system [47]. It provides sender and receiver anonymity. Mixminion uses TLS [34] which provides end-to-end perfect forward secure channel between sender and receiver. The user email passes through a cascades of mixes which involves permutations and delay before it reaches its recipient. Every mix server can see only its direct adjacent servers as it is customary in Mix-Net-based schemes.

Mixminion uses only a single-use reply blocks (SURBs) as a control against tracing the recipient path vulnerability of previous remailer types. To prevent *tagging* attack it utilizes a cryptographic checksum. In *tagging* attack, and active adversary modifies the message in such a way to have a tag which allows the adversary to trace it along the path.

In general, Remailer are high latency Mix-Net-based systems, thus our scheme have similar comparisons with them as that mentioned in section 2.1.1. Additionally, all the messages in our schemes are equiprobable which precludes tagging attack.

2.2 Communication-Based Anonymous Communication Schemes

In this section, we will study anonymous communication schemes that are not based on using cryptographic primitive. Although they may use cryptography for some parts of the anonymity solution, it does not constitute the major part. They replace computational-cost of using cryptographic primitive by the communication-cost resulted from using random routing, multicast, or broadcast transmission. Our secure k-anonymous query scheme, presented in Chapter 3, falls under this class of anonymous communication schemes. As we will show, the main problem here is the trade-off between incurred communication-cost and offered anonymity.

2.2.1 DC-Net Scheme

Chaum proposed DC-Net in 1988 [21]. The DC-Net achieves anonymity by embedding the message into a broadcast of other users' broadcast transmission. Although the DC-Net name comes from the problem of dining cryptographers introduced by Chaum, the techniques itself is essentially based on broadcast transmission of users' messages where the communication-cost is quadratic in the number of participating users.

The problem behind DC-Net design is that a group of cryptographers get paid for their dinner anonymously by either the National Security Agency (NSA), or one of them. They are curious to know if they get paid by the NSA. The DC-Net offers a scheme to solve the problem by having each cryptographer flips a coin, and exchange the result with her neighbor on the right hand side. Then all of them broadcast the XOR of result of the two flips they know. That is for user i, she knows her flip result (f_i) and her neighbor's flip result f_j, then she will broadcast ($f_i \oplus f_j$). However if any one of the cryptographers has paid for their dinner, she must do additional XOR with 1 before broadcasting the result. The cryptographers will know if any one of them pays for their dinner when the modulo-2 sum (XOR) of the broadcasted messages is 1, otherwise the NSA is the one who pays for their dinner. This is true because they constitute a circular graph, and therefore their messages will have similar terms that cancel each other except for the 1 that is added by the cryptographer who paid for their dinner. This protocol sends a single bit anonymously, and similar approach is used to send messages with arbitrary length.

The main issues with DC-Net is that it incurs bandwidth overhead because it utilizes broadcast transmission. The communication-cost incurred to send a single bit anonymously is in the order of $n(n-1)$, where n is the number of users. That is, the communication-cost is quadratic in the number of participants. This is because DC-Net trades high communication-cost for maximum level-of-anonymity that is equal to the size of network n. Although DC-Net does not scale well with the network size n, it is immune to traffic analysis attacks such as intersection and statistical disclosure attacks since it uses a single anonymity-set of size n. DC-Net provides an information-theoretic perfect secrecy solution that is analogous to One-Time Pad in cryptography.

The main attacks against DC-Net are users collusion and message collision attacks. Users can collude to reveal the identity of the sender. In addition, an insider attacker can choose to send a message every run of the DC-Net protocol, or just drop the protocol causing denial of service attack. Ways to prevent message collision attack are based on the idea of assuring that at any one time a single user is allowed to send a message in the broadcast transmission which prevent collisions. This is achieved by different methods in the literature such as having a *reservation* protocol to reserve one slot for each user has message to send [48], or using Xor-trees [49] which prevents two user from sending a message in the same time.

Our secure k-anonymous scheme is similar to DC-Net since they are both provably secure, low latency communication-based schemes that suffers from the consequences of the trade-off between incurred communication-cost and offered level-of-anonymity. However, our scheme is provably secure, much simpler, and scalable. It achieves most of its scalability by allowing the level-of-anonymity k to vary from 1 to n, where n is the size of network, instead of offering maximum anonymity ($k = n$) at all times, as in the case of DC-Net. Moreover, our trust-none model provides a proactive solution against collusion attacks since the client trusts no other entity in the scheme, hence the security of the scheme does not rely on the honesty of other entities.

2.2.2 Crowds Scheme

Crowds scheme is designed to provide anonymity for users browsing the Internet without using a cryptographic primitive [50]. It basically replaces the computational-cost of cryptography by the communication-cost of additional random forwarding within a crowd of users. The user first joins a crowd which is a list of other Internet users that can be downloaded from a central server. When she wants to connect to a Web server, she forwards her web request to a randomly selected member of the crowd. The recipient tosses a coin to decide on either sending the Web request directly to the Web server, or keep forwarding it to another member of the crowd in which case the process is repeated. The coin is biased towards forwarding through the crowd member.

Each member of the crowd knows its adjacent members only, and there is no way to classify the previous crowd member as an initiator of the Web request or a forwarder. Hence Crowds scheme achieves *source anonymity* against local attacker who is allowed to control part of the crowd and/or the end Web server. However since the Web request and response are forwarded in a clear text using the same path, the anonymity is no longer maintained after the first compromised member of the crowd [33].

The values of probabilities of forwarding to another crowd member and sending to the server affect the trade-off between communication-cost and achieved anonymity. Larger forwarding probability value tends to increase the the communication-cost and anonymity, while a small value decreases the achieved anonymity and communication-cost [50].

Leaving the perfectly secure yet costly maximum anonymity environment of DC-Net, Crowds scheme is vulnerable to traffic analysis attacks such as intersection and predecessor attacks [39, 40]. In this statistical attack, the adversary counts the number of times a Web resource is accessed by the same crowed member, and identify the one with the most request as the actual source.

Our scheme provides k-anonymity as further enhancement to the communication-based approach in order to tackle the problem of trade-off between incurred communication-cost and offered anonymity. On the other hand, Crowds scheme utilizes a random walk strategy to achieve the same goal considering large scale network. In addition, our scheme adopts disjoint anonymity-sets as a countermeasure against intersection and statistical attacks.

2.2.3 K-Anonymous Schemes

Our scheme falls under this subcategory of communication-based class of anonymous communication schemes. The k-anonymous schemes come as an application of k-anonymity [51] in the communication-based anonymous communication schemes to tackle the problem of trade-off between communication-cost and anonymity. In theses schemes, the sendr/receiver of a message is indistingushable from others with a probability non-negligibly greater than $1/k$, where k is the offered level-of-anonymity.

To solve the problem of scalability in DC-Net scheme, Herbivore [52], CliqueNet [53], Ahn et al. [7], Carbunar et al. [8], and De Cristofaro et al. [54] divide the network of size n into multiple DC-Nets, called cliques, groups or regions that serve as anonymity-sets, of size $k \in [1, n]$. These schemes are vulnerable to traffic analysis attacks by global adversary including intersection and statistical attacks since they allow k to vary below the maximum anonymity of $k = n$ that is achieved by DC-Net. That is, an adversary, who intercepts the scheme for sufficiently long time, may assign different probability values to different members of a clique or group to reveal the victim identity.

As a step towards addressing such risks, Herbivore allows nodes to launch intersection attack on themselves as a self penetration testing step to check whether the offered anonymity has been compromised. Additionally, they provide another protection measure by introducing Exit phase to limit the changes to clique size during long-running anonymous communications.

To defend against global eavesdropper, the work of Carbunar et al. [8], presents an anonymous query scheme that utilizes several k-anonymous transformations namely: the Union, Randomized, and Hybrid transformations, to anonymize a sequence of queries Q in the context of WSN. Each query corresponds to a destination node in the WSN. These transformations produce a larger sequence of destination nodes M than the sequence of queried nodes at their input Q by adding additional nodes. That is, each query $q_i \in Q$ is transformed into a set of WSN nodes. Compared to our scheme which transforms each individual query into its corresponding anonymity-set of size $k \in [1, n]$, this scheme introduces more latencies in the system by requiring to store a sequence of queries before execution. Moreover, since our attacker model allows the adversary to know the whole anonymity scheme including the generated disjoint anonymity-sets (see section 1.2.3), we gain no security advantage from anonymizing a sequence of queries.

The work of De Cristofaro et al. [54], which can be considered as a modification or an extension of Carbunar et al. [8], introduces a k-anonymous querying protocol for WSN environment that collects response from all nodes along the source route in a piggyback manner. That is each node in the source route attaches its response to the single query packet. Being k-anonymous schemes, these schemes are vulnerable to intersection and statistical attacks.

Carbunar et al. [8] k-anonymous scheme applies keyed-Hash Message Authentication Code (HMAC) on the client query to make it indistinguishable from a random string. This defends against a local eavesdropper who have access to only part of the WSN, such as the servers that connects the client to the WSN. We argue that the HMAC provides a confidentiality service that fails to protect against a global eavesdropper. The global eavesdropper have access to the whole WSN, and able to gather information about the route of the query and response path. As countermeasures against such global attackers, they present several anonymity transformations that query dummy destinations in addition to the true one. De Cristofaro et al. [54] scheme uses onion routing to enhance the anonymity obtained from collecting responses from all hops along the source route to and from the true destination. They consider local attacker who is able to intercept intercommunication among only infected nodes; the nodes under the control of the attacker. The maximum anonymity is achieved when the true destination is indistinguishable from all of WSN nodes, that is $k = n$. Since onion routing encrypts the intercommunication traffic among WSN nodes, and their attacker model is local, then the maximum anonymity is achieved when the number of infected nodes is zero. Nevertheless, the anonymity degrades to the number of source routes hops when all the WSN nodes are infected. Notwithstanding the encryption of traffic provided by the onion routing, if a global eavesdropper is contemplated, this scheme can not offer the true destination better anonymity than the number of the source route hops. It becomes obvious that the encoding of the network traffic using HMAC or onion routing does not have any advantage on the offered anonymity level when a global eavesdropper is adopted. As a deduction, we take a radically different approach that adopts no encoding techniques. Instead, we allow our powerful global eavesdropper to see the contents of the source route without sacrificing the offered anonymity. In short, we decouple security services, such as confidentiality, Integrity, and authentication, from anonymity service to construct a pure anonymity solution. These security services is customarily provided using standard cryptographic techniques which is out of the scope of this work.

Both schemes [8, 54] utilize different cryptographic algorithms such as ECC, AES, HMAC, and MD5 whereas we are using none of them for the reasons explained above.

This results in a significant saving in the incurred computational-cost, delay and storage requirements of the keys that makes our approach a more suitable one for a scarce resources environments such as that of WSN [55], sensor-cloud-based IoT systems, and low latency Web and real time applications.

Furthermore, both of Carbunar et al. [8] and De Cristofaro et al. [54] schemes maintain constant size routing packet using techniques of padding or appending the routing hops. While this conceals the position of the current routing hop from local eavesdropper, we find that, when a global eavesdropper is considered, it is unnecessarily increases the exact communication-cost by a factor of $(\frac{2}{1+\delta})$, where δ is the inverse of the length of the routing path[1]. However, asymptotically our work and theirs incur the same.

Our trust model is also different from that of Carbunar et al. [8] and De Cristofaro et al. [54]. Colluding of the two servers of Carbunar et al. [8] makes them as powerful as a global eavesdropper. Similarly, colluding of WSN nodes in De Cristofaro et al. [54] results in the same. alternatively, we propose an effective mechanism to anonymously querying the WSN in an environment where the anonymity beneficiary (the client) trusts no party other than herself. Thus, the attained anonymity does not depend on how honest other system components behave.

Most of these schemes, even in their simulation results, have not analyzed the trade-off between query-anonymity and incurred communication-cost along the whole range of anonymity level. This is mainly because of the possibility of the intersection and statistical attacks which varies in its severity based on the client access pattern. Since this access pattern is modeled as random process, hence the level-of-anonymity is not deterministic. For

[1] Let x denotes the source route path length, then the communication-cost of the constant size source route packet is equal to $x + x + \cdots + x = x \times x = x^2$. Whereas, the communication-cost of the variable size source route packet is $x + (x-1) + \cdots + 1 = \frac{x(x+1)}{2}$. Thus the cost ratio is $\frac{x^2}{\frac{x(x+1)}{2}} = \frac{2}{1+\delta}$, where $\delta = \frac{1}{x}$.

example, De Cristofaro et al. [54] provide a WSN querying protocol without characterizing the trade-off between its achievable query-anonymity and its incurred communication-cost. They measure the cost practically and independently from the obtained level-of-anonymity. As a step towards addressing this lack in the literature, the cost-anonymity trade-off analysis (Chapter 5) that utilizes our secure query k-anonymity notion (Chapter 3) is the central focus of this work.

2.3 Anonymity Notions and Metrics

Chaums introduced the notion of anonymity-set s in the DC-Net [21] in 1988. The well-cited definition of anonymity-set is given by [56]: "Anonymity of a subject means that the subject is not identifiable within a set of subjects, the anonymity-set."

Since then, different anonymous communication solutions and protocols have been proposed. Additionally, a variety of countermeasures have been investigated to protect against possible attacks that compromise the anonymity solutions. However, most of these systems assume that the attacker can do no better than a assigning equally likely a posteriori probabilities of $1/|s|$ to the anonymity-set members, where $|s|$ is the size of anonymity-set. From the information theoretic viewpoint, this means the attacker learns no information from intercepting the anonymous communication scheme [16]. Thus, anonymity-set size $|s|$ comes in handy as a measure to evaluate the quality of anonymity.

Berthold et al. [36] uses a different anonymity measure which is equal to $log_2|s|$.Yet the equally likely a posteriori probabilities environment is also taken for granted for this measure to work because $log_2|s|$ measures the information needed by attacker to break the anonymity system when the a posteriori probabilities are equally likely as it is shown by [57, 58].

Analogous to Berthold et al. [36], Herbivore [52], CliqueNet [53], Ahn et al. [7], Carbunar et al. [8], and De Cristofaro et al. [54] specify no mechanism to fully control the information leak vulnerability, and prevent traffic analysis attacks such as intersection or statistical attacks. Thus, there were no clear measurement model for the quality of offered anonymity. In their performance analysis, Herbivore [52] uses the clique size which is equivalent to the anonymity-set size as a performance index without satisfying the condition of equiprobable a posteriori probabilities. Ahn et al. [7] provide no implementation, and thus no performance analysis for their scheme.

Realizing that intra-redundancy may occur within Q, the sequence of queries inputted to the anonymity scheme, and/or within output sequence M, the notion of anonymity of Carbunar et al. [8] considers only the unique WSN nodes in Q and M, which are denoted as $\tilde{Q}$ and $\tilde{M}$ respectively. $\tilde{M}$ set is basically the anonymity-set. Thus they define query-anonymity $S(T)$ (called "*spacial privacy*" there) as the inverse of the probability of guessing the true destination given $\tilde{M}$, specifically $S(T) = |\tilde{M}|/|\tilde{Q}|$. The maximum achievable anonymity arises when the true destination is indistinguishable from the rest of WSN nodes of size n. For example, considering a $\sqrt{n} \times \sqrt{n}$ square grid WSN, when the input query sequence length is 1 (i.e. $|\tilde{Q}| = 1$), the maximum anonymity $S(T) = n$ is achieved when $|\tilde{M}| = n$. That is when all of WSN nodes appears in the output sequence as possible destinations. On the the contrary, when $|\tilde{M}| = 1$, no anonymity is provided since $S(T) = 1$. Carbunar et al. [8] uses the *temporal privacy* notion to refer to the concealment of query frequency information.

Although not mentioned, the implicit assumption for their query-anonymity definition [8] to be useful is that the events of M members being true destinations are mutually exclusive events with equally likely probabilities. This is exactly the same conditions under which the concept of anonymity-set operates [59, 60, 61]. However their scheme offers no control on the inter-redundancy or intersection among different output sequences. This leaks information to the adversary, and opens doors to intersection and statistical attacks that breaks the equiprobable probabilities requirement [62, 63, 64]. In intersection attack, as mentioned in section 2.1.2, the attacker intersects the output sequences of multiple runs of the querying protocol to narrow his search for the true destination. Depending on its severity, the intersection attack shrinks the level-of-anonymity to as low as zero which takes place when the intersection results in one destination node [65]. These observations make the above anonymity measurement model not a valid one since no level-of-anonymity can be claimed if the scheme is vulnerable to such attack. It is worth noting that their Dynamic Transform, which tries to uniformize the the probability distribution of M, provides a reactive solution to this problem even though this is not mentioned explicitly.

Reiter and Rubin in their work on Crowds[50], were the first to realize the need for the anonymity measurement model to consider the not-equally likely a posteriori probabilities that is assigned by attacker to the anonymity-set members. In other words, the equally likely a posteriori probabilities is not always guaranteed. However, their probabilistic anonymity metric gauges anonymity of each member of the anonymity-set separately instead of the whole anonymity-set. Consequently, it does not measure the quality of anonymity for the entire anonymity sheme [33]. Later, this approach is referred to as *local* anonymity measure [66], in opposite to *global* anonymity measure which uses a system-wide metrics. In Stop-and-Go-Mix, Kesdogan et al.[28] acknowledges the possibility of not-equally likely a posteriori probabilities, nevertheless it is not investigated further.

In their important step to standardize the anonymity terminology, Pfitzmann and Kohntopp [67, 56] put forward a simple scale for anonymity measurement by stating that, “ Anonymity is the stronger, the larger the respective anonymity-set is and the more evenly distributed the sending or receiving, respectively, of the subjects within that set is”. Thus, they affirm that the probability distribution of the anonymity-set members has a significant impact on the quality of anonymity measurement in addition to the anonymity-set size. Yet no mathematical measurement model is articulated to determine the quality of anonymity.

Inspired by the concept of a posteriori probability distribution of anonymity-set members that is introduced by Pfitzmann and Kohntopp [67], Serjantov and Danezis [58] uses the Shannon entropy [68] as a system-wide metric to measure the additional average information per anonymity-set member that attacker needs to identify the actual sender/receiver of a message. Their measurement metric is called *Effective Anonymity-Set Size* (S) which is defined as, $S = H(X) = -\sum_{i=1}^{|s|} P(x_i) log_2(P(x_i))$, where X is a discrete random variable that denotes the actual sender/receiver of a message within an anonymity-set $\{x_1, \ldots, x_{|s|}\}$, $P(x_i)$ is the a posteriori probability value that is assigned by the attacker to each member x_i of the anonymity-set, and $|s|$ is the actual anonymity-set size. The main issue with the *Effective Anonymity-Set Size* metric is that it fails to calculate the *Effective Anonymity-Set Size* as equal to actual anonymity-set size $|s|$ for the equally likely case, i. e. $P(x_i) = \frac{1}{|s|}$.

Independently, Diaz et al. [57] suggests a similar *global* anonymity measure to normalize the above entropy $H(X)$ by the maximum entropy $H_{max} = log_2|s|$, which occurs when the a posteriori probabilities of the anonymity-set members are equally likely. They call it the *Degree of Anonymity* $d = \frac{H(X)}{H_{max}}$. This metric solves a problem with the *Effective Anonymity-Set Size* proposed by Serjantove and Danezis [58] which is lacking of a measurement reference point. However, unlike Serjantov and danezis, this metric ignores totally the role of the anonymity-set size. For example, when the a posteriori probabilities are equally likely, $H(X) = H_{max}$, thus *Degree of Anonymity* $d = 1$ regardless of the anonymity-set size $|s|$.

A number of information-theoretic efforts have followed Diaz et al. and Serjantov et al. [57, 58]. Tóth et al. [66] uses a *local* anonymity metric that is based on maximal a posteriori probability, which corresponds to the min-entropy, of each anonymity-set member being compromised. Their main argument is that Shannon entropy measures the average not the worst-case for each anonymity-set member. Clauß et al. [69] invetigate a generalization for Shannon entropy, min-entropy, and max-entropy. A particular entropy-based approaches are taken to use relative entropy in [70] and mutual information in [71, 72] to compare the a priori and a posteriori knowledeg of the attacker before and after interception of the anonymity scheme, and measure the information leakage for different traces of the anonymity scheme.

Realizing that having a non-uniform probability distribution of the anonymity-set members is a vulnerability, which opens the door for a variety of threats, stimulates research different traffic analysis attack scenarios namely, intersection, and predecessor or statistical disclosure attacks [64, 73], as discussed in section 2.1. These attacks aims at exploiting the leakage of information due to the non-uniform a posteriori probability distribution.

The novelty of our work stems form that we are studying the problem from a totally new point of view. Instead of studying what happened to the anonymous communication system under not-equally likely probabilities of anonymity-set members environment, we scrutinize the enabling conditions that establish the equally likely case. In addition we shed a light on how to design and implement a practical solution that guarantees a specific level-of-anonymity, which is equal to the anonymity-set size, under unconditional security.

These observations prompted us to adopt a unique approach that avoids intersection attack by using disjoint anonymity-sets with distinct members [14, 17, 18, 19] .This proactive approach has been stated briefly by De Cristofaro et al. [54] as follows,“we can let the client select the same route for the same queried node. Thus, ADV can not intersect its views of query executions", (ADV is the adversary). Nevertheless, they did not mention how to accomplish that since they require all of hops along the source route to respond to the client's query. In this work, we maintain the disjoint property of the anonymity-sets by having the WSN nodes that connect the client to an anonymity-set forward the query but do not respond to it. To the best of our knowledge this is the first preventive measure to defend against intersection and statistical attacks.

Moreover, in contrast to [52, 53, 8, 9] that avoid defining anonymous channels, our secure query k-anonymity notion defines, in an application agnostic manner, the fundamental properties of an arbitrary querying protocol to be successful in achieving a specific level-of-anonymity. Our indistinguishability-based query-anonymity definition has certain commonalities with [74, 75, 76, 77, 78] in that it follows analogous approaches to what is used in the formalization of semantically secure encryption schemes [79, 80, 81] to define the anonymity notion. In addition, our information-theoretic framework shares some commonalities with [61, 60, 66] in that they all adopt Shannon's information-theoretic approach [16] to formalize the notion of anonymity. Yet, our notion stands out in several aspects.

1. Our notion provides an indistinguisahbility-based characterization of k-anonymity [7, 51] without compromising their standard meaning in the literature [59].

2. It bridges the gap between indistinguishability-based formulation of anonymity notions and the standard, and intuitive concept of anonymity-set in a natural way. These two anonymity characterizations are treated as alternatives in the literature [74].

3. It can be related naturally with other security notions namely, Shannon's *Perfect Secrecy* [16].

4. It extends itself naturally to define the meaning of “secure” for query k-anonymity.

5. It also proves to be beneficial in formulating the design principles of anonymity constructions that maintain secure query k-anonymity in the presence of eavesdropping adversary who intercepts the system for sufficiently long time.

Indeed the absence of a query-anonymity definition that captures the practical concerns of anonymous communication systems motivated this work. The query-anonymity definition proposed in this work is proved to be helpful in the design, analysis of trade-off between query-anonymity and communication-cost, and evaluation of provably secure anonymous communication schemes that are immune to intersection attack. We provide a summary of the comparison of our work with existing K-Anonymous Communication Schemes in Table 2.1.

Table 2.1: Comparison with Existing k-Anonymous Communication Schemes

Criterion	**Related Work**	**Our Work**
Trade-off analysis	None or partail	✓
Disjoint anonymity-Sets	X	✓
Traffic analysis attack	Vulnerable	Immune
Anonymity metric	Probabilistic	deterministic
Secure k-anonymity def.	X	✓
Cost-Benefit analysis	X	✓
Location anonymity metrics	X	✓
Attacker Model	Mostly Passive	Unconditional Passive
Global/Local	May differ	same power
Collusion/Proxy	some are vulnerable	X

Chapter 3

A Theoretical Framework for Secure K-Anonymous Query Schemes

We consider k-anonymous query schemes in sensor-cloud-based IoT systems which are k-anonymous communication schemes that provide receiver-anonymity for their clients (see section 2.2.3). One of the first challenges that we tackle when we seek to analyze the trade-off between query-anonymity and communication-cost is: how do we define “secure” in the context of k-anonymous query scheme? In other words, what are the essential characteristics of a secure k-anonymous query scheme? Then, how do we design a secure k-anonymous query scheme which never become vulnerable in the presence of eavesdropping adversary with unlimited time and computational power? How immune is a secure k-anonymous query scheme to traffic analysis attack in which single or multiple outputs are intercepted?

Defining the security of secrecy schemes formally is of special importance because these schemes are designed to operate naturally in malicious environment where adversaries try all the possible ways to break them. Under the assumption of some attacker model, definition of security of a scheme provides a precise idea about what the secrecy scheme must achieve to prevent the adversary from mounting a successful attack against it. Proof of security of an instance of a security scheme determines the level at which it satisfies the definition of security. Furthermore, defining what it means for a secrecy scheme to be secure, helps in the comparison and evaluation of various security schemes according to their degree of satisfiability to the definition of security [81].

In this chapter, we provide a formal definition of security of k-anonymous query scheme building on two well-known security models used for encryption namely, the ciphertext indistinguishability under chosen plaintext attack (IND-CPA), and information-theoretic notion of perfect secrecy [82, 80, 81, 16]. We further shed a light on how to design and implement a practical solution that guarantees a specific level-of-anonymity under unconditional security. In Chapter 4, we analyze the Disjoint Anonymity-Sets scheme (DAS) to prove that it is secure according to our definition of security.

3.1 Notations of k-Anonymous Query Scheme

In this section, we give general notations of a k-anonymous query scheme in the considered sensor-cloud environment. Let $\mathbb{Q}$ denote the set of all possible queries, where each $q \in \mathbb{Q}$ is associated with the destination node being queried. Let every node be a valid destination of a query, i.e., $|\mathbb{Q}| = n$. For simplicity, a reliable communication channel between each pair of adjacent nodes is available all the time. Dropped messages (queries or responses) due to bad channels may affect the offered level-of-anonymity of a k-anonymous query, which is left for future research.

The k-anonymous query scheme provides receiver anonymity service by hiding the ID of the destination node of a query in the crowd of other $k-1$ queries' destinations [51, 7, 52, 53, 8, 9]. This is achieved by querying additional $k-1$ nodes along with the original query, such that an adversary cannot learn the true destination of a k-anonymous query with a probability non-negligibly greater than $1/k$ by analyzing its traffic patterns [7, 8].

To simplify the exposition, we assume that each query $q \in \mathbb{Q}$ is sent to one destination node $v_d \in V$ [8, 14], where WSN is modeled as a graph $G(V, E)$ (see section 1.2.2). This allows us to define a one-to-one correspondence that uniquely maps each query in the query space to a specific destination node. That is, the destination node of a query $q \in \mathbb{Q}$ is given by the function $f : \mathbb{Q} \mapsto V$. We define the value of each query and its response abstractly to be equal to the ID number of its destination node. We refer to a query and its response by the same symbol q. For example, q_1 is a query/response that corresponds to destination node v_1 whose ID number is 1, and so on. We justify this convention by considering the fact that the main function of k-anonymous query scheme is to conceal the ID number of destination node of a query/response regardless of its data content. In other words, the only sensitive information is the ID number of the destination node that is being queried or responding to a query. Furthermore, the value of k-anonymous query is the ID of its *true* destination node.

For $q \in \mathbb{Q}$, $Pr[Q = q]$ denotes the probability that the true destination of the k-anonymous query is q, where Q is a discrete random variable denoting the query value. This is used to model the fact that an adversary who is conducting traffic analysis against the k-anonymous query scheme may assign different probabilities to multiple query values of different destinations based on her prior knowledge. As an example, for a 2-anonymous query ($k = 2$) with $\{q_1, q_2\}$, an adversary may know somehow that $Pr[Q = q1] = 0.4$, and $Pr[Q = q2] = 0.6$.

3.2 A General k-Anonymous Query Construction

A generic k-anonymous query scheme, denoted by (π, T, T^{-1}), is made up of one partition algorithm and two anonymity transformations, which are defined as follows.

1. π is the partition algorithm which takes as input the whole WSN of size n destination nodes, and constructs anonymity-sets each of size $|s| \in [k, 2k-1]$ destination nodes that is required by the anonymity transformation T, where $k \in [1, n]$ is the level-of-anonymity. The case $|s| = 1$ is when no k-anonymous query can be offered. To provide *uniform anonymity*, the size of the anonymity-sets should be a system-wide constant. For a fixed n and k, let $\mathbb{S}$ denotes the set of all possible anonymity-sets. Recall that the query value is the ID of the corresponding destination node with a one-to-one correspondence, each anonymity-set $s \in \mathbb{S}$ is substantially a set of queries associated with a set of destination nodes of size $|s|$. For $s \in \mathbb{S}$, let $Pr[S = s]$ denote the probability that the anonymity-set is s, where S is a discrete random variable taking on the set of possible values of anonymity-sets $\mathbb{S}$. The partition algorithm may be probabilistic so that it might result in a different set $\mathbb{S}$ every time it runs.

2. $T : Q_T \to \mathbb{S}$ is the anonymity transformation of one space (i.e., the set of possible queries Q_T) to a second space (i.e., the set of possible anonymity-sets $\mathbb{S}$). We assume that queries in $\mathbb{Q}$ are fully partitioned into multiple sets of Q_T, denoted by $\mathbb{Q}_{\mathbb{T}}$, in such a way that every query is included in one and only one Q_T. In other words, $\mathbb{Q}_{\mathbb{T}}$ are disjoint sets. Thus, T maps a query $q \in Q_T$, which is called true-destination query, to an anonymity-set $s \in \mathbb{S}$, such that $s = \{q, S_b\}$. Hence, for every $q \in Q_T$, it holds that: $s = T(q) \to q \in s$. S_b is a set of bogus-destination queries as opposed to the true-destination query q, i.e. $q \notin S_b$. An important technicality is that there

is no limitations on the definition of the set of bogus-destination queries S_b as long as S_b is a proper subset of $\mathbb{Q}$, i.e. $S_b \subset \mathbb{Q}$. Based on the affordable cost, s can be as large as $\mathbb{Q}$ in which case maximum level-of-anonymity $k = n$ is provided. That is, $s \subseteq \mathbb{Q}$. Hence the size of anonymity set $|s| \in \{1, 2, 3, \ldots, n\}$, where $n = |\mathbb{Q}|$ is the size of WSN.

3. $T^{-1} : \mathbb{S} \rightarrow Q_T$ is the inverse transformation of T that maps an anonymity-set s to the ID number of its true-destination node q. We assume T^{-1} is a deterministic procedure that always succeeds. That is $T^{-1}(s) = T^{-1}(T(q)) = q$.

The transformation T provides anonymity through the concealment of a true-destination query q within S_b. We stress that, considering a passive eavesdropping adversary $\mathcal{A}$ who sees a single anonymity-set s, for every $q \in s$ to be contributing to anonymity with a none-zero value, it holds that: $q \in s \rightarrow (0 < Pr[Q = q] < 1)$. A special case is when $Pr[Q = q] = 0$, which refers to a bogus-destination query that the adversary throws with no effort. Thus, $Pr[Q = q] = 0$ is assigned for any $q \notin s$ if we assume that adversary sees a single anonymity-set.

To use the described anonymity scheme, the partition algorithm π is first run to generate anonymity-sets $\mathbb{S}$, in which each set $s \in \mathbb{S}$ is of size $|s| \in [k, 2k - 1]$. A query $q \in Q_T$ is selected by the client and this choice information is stored. The particular transformation T corresponding to Q_T is applied to the selected query q to produce an anonymity-set $s \in \mathbb{S}$. This anonymity-set s of the destination nodes is queried by a reliable channel that may be intercepted by an adversary. In addition, the adversary is able to intercept the responses that client collects from anonymity-set s. Finally, the client applies the inverse transformation T^{-1} to the anonymity-set s in order to recover the true destination ID number, and read its data. Basically, the client drops every response she receives except from the true destination.

Note that the adversary *a priori* information/knowledge is represented by the *a priori* probabilities associated with each possible query $q \in Q_T$ to be the true-destination query of k-anonymous query, *before* observing the anonymity scheme running. On the other hand, the *a posteriori* probabilities of possible queries in Q_T to be true-destination query, which is calculated by the adversary *after* observing the anonymity scheme running, constitute the adversary *a posteriori* information/knowledge.

We now investigate the conditions under which the k-anonymous query scheme is secure in the sense that it leaks no additional information to the adversary [81]. Towards that, we introduce an information-theoretic perfect secrecy [16], and adversarial-indistinguishability [82] definitions of secure query k-anonymity in the presence of eavesdropping adversary armed with two types of traffic analysis capabilities namely, interception of single anonymity-set and interception of multiple anonymity-sets.

3.3 Interception of single anonymity-set

In this section, we propose two equivalent but powerful definitions of secure query k-anonymity. Towards the end, we investigate the practical aspects of k-anonymous query scheme which achieves secure query k-anonymity in the presence of unconditional eavesdropping adversary.

3.3.1 Information-Theoretic Perfect Secrecy Approach

We start by proposing a definition for query $k - anonymity$.

Definition 1 (query $k-anonymity$)**.** *An anonymity scheme* (π, T, T^{-1}) *achieves query* $k-anonymity$ *against eavesdropping adversary* $\mathcal{A}$ *if* $\forall \mathcal{A}$, $\forall q_j \in Q_T$, *where* Q_T *constitutes the input space of* T, $\forall s \in \mathbb{S}$ *that are resulted from* $s = T(q_j)$, *there exists a subset of unique queries* $Q_k = \{q_1, \ldots, q_j, \ldots, q_k\}$ *where* $Q_k \subseteq Q_T$, $k = |Q_k| \in [1, |Q_T|]$, *we call the subset of indistinguishable queries* Q_k, *such that the calculated probabilities by the adversary over* Q_k *during traffic analysis (the a posteriori probabilities) are equally likely,*

$$Pr[Q = q_i \mid S = s] = \frac{1}{k}, \ \forall q_i \in Q_k \tag{3.1}$$

It is easy to see that the case of $k = 1$ offers no anonymity, while in the case of $k = |Q_T|$, the *a posteriori* probabilities that $\mathcal{A}$ associates with each of Q_T's members being the true-destination query are equally likely.

Now, the *apriori* information of $\mathcal{A}$ comes merely from the fact that the adversary is allowed to learn the details of the anonymity scheme by the *security by transparency* requirement [16, 83]. This adversary prior information is that the true-destination query is chosen at random out of Q_T, the input space of anonymity transformation T. Thus, the *apriori* probability distribution over Q_T is uniform,

$$Pr[Q = q_i] = \frac{1}{|Q_T|}, \ \forall q_i \in Q_T \tag{3.2}$$

If the anonymity scheme is so secure that it leaks no additional information to the adversary when it is executed, the adversary information stays at its prior level. Hence, adversary *a posteriori* knowledge represented by *a posteriori* probabilities are the same as *a priori* knowledge that is represented by *a priori* probabilities. As indicated before, this point acts as the maximum level-of-anonymity ($k = |Q_T|$ and $Q_k = Q_T$) that can be obtained from the anonymity scheme. By substituting the value of $k = |Q_T|$ in *a posteriori* probabilities equation of query $k - anonymity$ (equation 3.1), *a posteriori* probabilities come to be equal to *a priori* probabilities as defined in equation 3.2. Now we have the position to provide an information-theoretic-based definition for the secure query $k - anonymity$ which is equivalent to the indistinguishability-based definition introduced in section 3.3.2:

Definition 2 (secure k-anonymous query scheme)**.** *An anonymity scheme* (π, T, T^{-1}) *achieves secure query* $k - anonymity$*, at which* $k = \min\limits_{Q_T \in \mathbb{Q}_\mathbb{T}} |Q_T|$*, and* $|Q_T| = \Theta(k)$*, against eavesdropping adversary* $\mathcal{A}$ *if and only if* $\forall \mathcal{A}$*,* $\forall q \in Q_T$*, where* Q_T *constitutes the input space of* T*,* $\forall s \in \mathbb{S}$ *that are resulted from* $s = T(q)$*, the calculated probabilities by the adversary over* Q_T *during traffic analysis (the a posteriori probabilities) have a uniform distribution, and are same as the a priori probabilities. That is,* $Pr[Q = q_i \mid S = s] = Pr[Q = q_i] = \frac{1}{|Q_T|}$*,* $\forall q_i \in Q_T$*.*

Stated another way, the necessary and sufficient condition to achieve secure query $k - anonymity$ is that the *a posteriori* probabilities are same as the *a priori* probabilities. Interestingly enough, using the information theory terminology [84, 61, 60], the entropy $H(Q)$ of *a posteriori* probabilities, which is a measure of uncertainty, is equal to the entropy of the *a priori* probabilities, and it is at its maximum value $H_{max}(Q) = \log_2 |Q_T|$ since the probabilities are equiprobable. That is since all possible events are equiprobable, the adversary uncertainty about the true-destination query is the highest. Another way of interpreting Definition 2 is that the probability distributions of true-destination query over the input space Q_T of the anonymity transformation T, and over its output space $\mathbb{S}$ are independent.

Remark 1. *For a desired level-of-anonymity* k*, the secure query* $k - anonymity$ *scheme* *is designed to have the subset of indistinguishable queries $Q_k \subseteq Q_T$ equal to its input space Q_T. That is $Q_k = Q_T$, and $|Q_T| = k$. However, when $\mathbb{Q}$ is partitioned into a set of Q_T, it is possible that n is not divisible by k. In this case, different $Q_T \in \mathbb{Q}$ will have different sizes specifically, $|Q_T| \in [k, 2k-1]$. Therefore, we have $|Q_T| = \Theta(k)$. In Definition 2 above, we determine k for a secure query $k-anonymity$ to be equal to the lowest size of $Q_T \in \mathbb{Q}_{\mathbb{T}}$. This is because $\forall Q_T \in \mathbb{Q}_{\mathbb{T}}$, it holds that: the maximum size of indistinguishable queries subset $Q_k \subseteq Q_T$ is equal to the lowest size of $Q_T \in \mathbb{Q}_{\mathbb{T}}$. This renders a compelling application of the concept of weakest link of the security chain [83, 85] in the context of anonymity. That is, the anonymity chain is only as strong as its weakest link.*

3.3.2 Adversarial-Indistinguishability Approach

To provide an indistinguishability-based definition of secure query k-anonymity, we use the following game with adversary $\mathcal{A}$, challenger $\mathcal{C}$, WSN of size n, and the anonymity scheme (π, T, T^{-1}). The game is the query-anonymity analogue to ciphertext indistinguishability under chosen plaintext attack (IND-CPA) property of cryptographic schemes [82, 81]. A cryptographic scheme possess IND-CPA property if the adversary fails to distinguish pairs of ciphertexts based on their corresponding pairs of plaintexts with a probability greater than random guess.

The game assumes that the partition algorithm has been run for a specific value of level-of-anonymity k on WSN of size n, and therefore the set of anonymity-sets $\mathbb{S}$ are constructed. We illustrate the game pictorially in Fig. 3.1.

The adversarial Indistinguishability game $G(\mathcal{A}, T)$

1. The adversary $\mathcal{A}$ chooses two arbitrary queries $q_i, q_j \in Q_T, q_i \neq q_j$, and send them to the challenger $\mathcal{C}$.

2. The challenger $\mathcal{C}$ replies with the anonymity transformation of one of the two queries denoted by q_b. That is, it replies with the anonymity set $s = T(q_b)$, after choosing a random bit $b \in_R \{i, j\}$, where $\in_R$ denotes choose at random.

3. The adversary $\mathcal{A}$ replies with $b' \in \{i, j\}$

4. The challenger $\mathcal{C}$ decides the adversary's success if $b' = b$

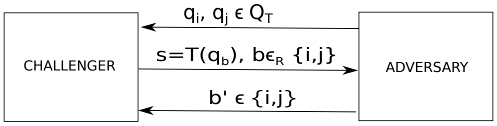


Figure 3.1: The adversarial Indistinguishability game $G(\mathcal{A}, T)$

We say that the anonymity scheme is secure or offers adversarial indistinguishability if the adversary's success is no better than a random guess. We are now ready for the formal definition.

Definition 3 (Perfect Indistinguishability)**.** *We say that an anonymity scheme* (π, T, T^{-1}) *is perfectly indistinguishable in the presence of unconditional eavesdropping adversary* $\mathcal{A}$ *if* $\forall\, \mathcal{A}$*,* $\forall\, q_i, q_j \in Q_T$*, where* Q_T *constitute the input space of* T*,* $\forall\, s \in \mathbb{S}$ *that are resulted from* $s = T(q_b)$ *as described in the game* $G(\mathcal{A}, T)$*, where* $b \in \{i, j\}$ *, it must hold that:*

$$Pr[Q = q_i \mid S = s] = Pr[Q = q_j \mid S = s] = \frac{1}{2}$$

The main goal of secure k-anonymous query anonymity scheme is to guarantee that an adversary can not learn the true destination of a k-anonymous query with a probability non-negligibly greater than $1/k$ by analyzing its traffic patterns [7, 8]. By proving Lemma 1 and Theorem 1 below, we seek to establish the relation between secure query k-anonymity and the perfect indistinguishability as stated in definition 3.

On the other hand, by intercepting the anonymity-sets, the adversary may gain additional information which can lower the level-of-anonymity offered by a vulnerable anonymity scheme. In the rest of this section, we will investigate the conditions under which the anonymity scheme is secure in the sense that it leaks no additional information to the adversary [16, 86, 81]. We seek to establish that by proving Lemma 1 and Theorem 2 below.

Lemma 1. *All $q \in Q_T$ that constitute the input space of a particular anonymity transformation T of a perfectly indistinguishable anonymity scheme (π, T, T^{-1}), as described by Definition 3, must belong to the same anonymity-set s, that is $Q_T \subseteq s$, and $|s| = |T(q)| \geq |Q_T|$.*

Proof. We prove by contradiction. If not all $q \in Q_T$ that constitute the input space of a particular anonymity transformation T of a perfectly indistinguishable anonymity scheme (π, T, T^{-1}) belong to the same anonymity-set, then there exist at least two queries q_i, q_j that belong to different anonymity-sets namely, $q_i \in s$, and $q_j \in s'$. Now when the game $G(\mathcal{A}, t)$ is executed, let us assume the adversary $\mathcal{A}$ chooses q_i, q_j and send them to the Challenger $\mathcal{C}$. It is obvious that $\mathcal{A}$ always succeeds in distinguishing the true destination of the anonymity-set that she receives from $\mathcal{C}$ with probability 1. For instance, when he receives s,

$Pr[Q = q_i \mid S = s] = 1$, $Pr[Q = q_j \mid S = s] = 0$. This contradicts Definition 3, and the result follows.

The result also follows from Definition 1. For the sake of contradiction, assume otherwise that not all $q \in Q_T$ belong to the same anonymity-set. Thus, in the eye of adversary who intercepts such anonymity-set, the size of subset of indistinguishable queries $|Q_k| < |Q_T|$. This is so because $Q_k \subseteq Q_T$ from Definition 1, and not all of Q_T appear in s

by assumption. That is, $k < |Q_T|$. This contradicts Definition 2 and Remark 1 of secure query $k - anonymity$ scheme. Hence all $q \in Q_T$ must belong to the same anonymity-set s. Consequently $Q_T \subseteq s$, and $|s| = |T(q)| \geq |Q_T|$.

□

Theorem 1. *A perfectly indistinguishable anonymity scheme* (π, T, T^{-1}) *achieves secure query k-anonymity with* $k = |Q_T|$.

Proof. The proof is straightforward once we run the game $G(\mathcal{A}, T)$ multiple of times that is sufficient to cover all the possible combinations resulted from choosing the two queries q_i and q_j arbitrarily out of Q_T. This entails that every $q \in Q_T$ is indistinguishable from the rest of queries in Q_T because of the pair-wise indistinguishability. Stated another way, the anonymity-set s of every $q \in Q_T$ which results from $s = T(q)$ contains no information about the true-destination query q. From lemma 1, all $q \in Q_T$ belongs to the same anonymity-set s. This permits every $q \in Q_T$ to act as a legitimate candidate of a true-destination query in the eye of the adversary who intercepts the anonymity-set s. Hence, to the adversary who observes the anonymity-set s, all $q \in Q_T$ are indistinguishable from each other. That is, the *a posteriori* probability distribution which is calculated by the adversary over Q_T is uniform,

$$Pr[Q = q_i \mid S = s] = \frac{1}{|Q_T|}, \; \forall \, q_i \in Q_T \tag{3.3}$$

Consequently a perfectly indistinguishable anonymity scheme (π, T, T^{-1}) achieves secure query k-anonymity, where $k = |Q_T|$.

□

An observation we can make from the proof of Lemma 1 above is that, an anonymity scheme that achieves a secure query $k-anonymity$ of level k provides all levels of anonymity that are less than k. This property is important when we consider WSN of size n that is not divisible by the level-of-anonymity k. We sate that formally in Corollary 1 below.

Corollary 1. *A secure query $k - anonymity$ scheme achieves all levels of anonymity $\leq k$.*

Proof. The results follows directly from Definition 2 of secure query $k-anonymity$ scheme which states that queries in Q_T are indistinguishable from each other. Thus, queries in any subset Q_s of Q_T are also indistinguishable from each other. That is, the achieved query $k - anonymity$ is of level $k = |Q_s| \in [1, \min\limits_{Q_T \in \mathbb{Q}_\mathbb{T}} |Q_T|]$ since $k = \min\limits_{Q_T \in \mathbb{Q}_\mathbb{T}} |Q_T|$ for secure query $k - anonymity$ scheme, as in Definition 2. Hence secure query $k - anonymity$ provides all levels of anonymity less than or equal to k.

□

We now consider the practical aspects of the anonymity scheme which achieves secure query $k - anonymity$. We basically subtract any incidental communication overhead cost without pulling down the offered level-of-anonymity, or contaminating the secrecy of the scheme. A simple consequence of Lemma 1 is the following.

Theorem 2. *To achieve secure query $k-anonymity$, it is sufficient to have the anonymity-set the same as the input space of the anonymity transformation, and the level-of-anonymity k equal to the size of the anonymity-set s. That is, $s = T(q) = Q_T$, and $k = |s| = |T(q)| = |Q_T|$.*

Proof. The result follows from Lemma 1, where for a secure query $k - anonymity$ scheme, we have $k = |Q_T|$, $Q_T \subseteq s$, and $|s| = |T(q)| \geq |Q_T|$. Furthermore, $k = |Q_T|$ is the maximum level-of-anonymity that a query $k - anonymity$ scheme can provide as stated in

Definition 1. This implies that all queries $q \in Q_T$ are indistinguishable from each other. Hence the queries in $s \setminus Q_T$, if any, are dummy queries contributing nothing to the level-of-anonymity. They just incur unnecessary communication overhead cost that we can avoid safely by having $s = T(q) = Q_T$. Thus, it is sufficient to have the anonymity-set the same as Q_T, and hence, the level-of-anonymity k equal to the size of the anonymity-set $|s|$, i.e $k = |s| = |T(q)| = |Q_T|$. □

We emphasize that in Theorem 2 and its proof we put no constraint on different anonymity-sets used to anonymize the same true-destination query. This is so because the adversary interception capability is restricted to a single anonymity-set.

3.4 Interception of multiple anonymity-sets

In this section, we investigate if the secure query $k-anonymity$, as formalized in Definition 2 for the single anonymity-set interception, is still guaranteed when multiple anonymity-sets interception is engaged. Furthermore, we formulate the design principles of anonymity constructions that maintain secure query $k - anonymity$ in such a vulnerable setting.

The capability of the adversary to intercept multiple anonymity-sets opens the door to more sophisticated attacks than just plain traffic analysis. The well-known types of such attacks are the *long-term intersection*, and *statistical disclosure* attack, [36, 37, 38, 35, 38, 64, 39, 40], as discussed in section 2.1.2. With the intersection attack, the adversary observes a large volume of network traffic in order to re-identify client's targeted destination. The adversary succeeded in breaking the anonymity scheme if the client maintains a consistent pattern of communications over time that leaks some information. By performing different non-empty intersection operations on the traced anonymity-sets, the adversary may gain information about the true anonymized value. Specific to our setting, this information leakage may result in non-uniform *a posteriori* probabilities of the different queries in the input space of the anonymity transformation namely, Q_T, leading to violation of the secure query $k - anonymity$ requirements as defined in Definition 2.

We make clear that the intent of this work is not to study the performance of the anonymity scheme that is vulnerable to intersection attack, but rather to investigate the design principles that help preventing the intersection attack. Note that the defence against intersection attack is an open problem that is difficult to solve efficiently [41, 42]. In particular, the disclosure attack is shown to be NP-complete. For example, the hitting set attack is one of the efforts for improving the computational efficiency of disclosure attack [87]. In turn, this complicates the measurement of anonymity under intersection attack, and makes the offered level-of-anonymity unpredictable. For that reason, prior work [52, 53, 7, 8, 9] which allows intersection attack avoids defining anonymous channel and measuring offered level-of-anonymity precisely, let alone analyzing the trade-off between query $k - anonymity$ and communication-cost along the whole range of level-of-anonymity.

As a prelude to determining the impact of intersection attack under our setting, we consider the following simplified example. Alice visited two hospitals for medication. A first rumor circulated in the press after Alice visited hospital A that she has either heart disease or flue, which we denote by anonymity-set $s_1 = \{heart, flue\}$. A second rumor circulated after Alice visited hospital B that she has either diabetes or heart disease, which we denote by anonymity-set $s_2 = \{diabetes, heart\}$. Since Alice is the *same* person about whom both rumors talk, a significant amount of information about Alice's true disease has been leaked to the *hones-but-curious* adversary. In fact the intersection attack between the two anonymity-sets in the example is so sever that it discloses Alic's real disease with high probability. It is easy to see that the intersection $s_1 \cap s_2 = \{heart\}$ results in one value for

Alice's real disease namely, heart disease. This in turn causes *a posteriori* probability that adversary assigns to heart disease to be higher than *a posteriori* probabilities she assigns to flue or diabetes. Thus heart disease is the most probable value of Alice's real disease. In this case, the *a posteriori* probabilities will be different from the *a priori* probabilities which have a uniform distribution intuitively. Hence the secure query $k-anonymity$ is sacrificed.

We point out that our setting is analogous to that of the above example in the sense that the query $k-anonymity$ scheme hides the interest of the *same* client who queries the WSN multiple of times. Therefore the intersection of multiple anonymity-sets means the same query (same destination node) appears in different anonymity-sets. Again, this makes it more probable than other queries to be the true-destination query, and thus violates the main property of secure query $k-anonymity$ namely, uniform *a posteriori* distribution.

It follows from the proceeding that to maintain secure query $k-anonymity$, the intersection among the multiple outputs of the anonymity transformation T traced by the adversary namely, multiple anonymity-sets, must be empty. We formalize this in Theorem 3 below.

Theorem 3. *Let* (π, T, T^{-1}) *denotes an anonymity scheme that achieves secure query* $k-anonymity$ *against eavesdropping adversary with the single anonymity-set interception capability. Let* $s = T(q)$ *denotes the resulted anonymity-set from anonymizing a query* $q \in Q_T$*, and let* Q_T *denote input space of anonymity transformation* T*. If eavesdropping adversary is capable of incorporating information gained from observing* $N_i \in [1, \infty]$ *runs of the anonymity scheme* (π, T, T^{-1})*, which resulted in* N_i *different anonymity-sets whose intersection contains the same query* q_i*, i.e. same destination node* $i \in [1, n]$ *in a WSN of* n *nodes, then the joint necessary and sufficient condition to preserve secure query* $k-anonymity$*, as defined in Definition 2, is that the anonymity-sets are disjoint and* $s = Q_T$*.*

Proof. Sufficiency: since the anonymity-sets are disjoint, it is impossible for the adversary who observes the anonymity scheme to assign a higher *a posteriori* probability to any of the destinations (query values) than others in the same anonymity-sets. Hence the *a posteriori* probabilities have a uniform distribution notwithstanding the number of intercepted anonymity-sets. It is easy to see that when $s = Q_T$, the *a posteriori* probabilities and the *a priori* probabilities are the same, and each of them equal to $\frac{1}{|Q_T|}$, which achieves secure query $k-anonymity$ as in Definition 2. It worth noting that, on its own, the first condition is not sufficient to provide secure query $k-anonymity$ for which we have to show that the *aposteriori* probabilities are the same as the *a priori* probabilities. The second condition namely, $s = Q_T$ jointly with the first one assures that.

Necessity: For the purpose of contradiction, assume otherwise that the intersection among anonymity-sets is not empty. Because the true-destination query of k-anonymous query reflects the interest of the same client, a query that is contained in the intersection of different anonymity-sets becomes more probable than others resulting in a non-uniform distribution for the *a posteriori* probabilities. This contradicts the requirements of the secure query $k - anonymity$ as in Definition 2. We claim that $s = Q_T$ is an implied necessary condition to have the intersection among different anonymity-sets empty.

To see why, we stress that $\mathbb{Q}$ is fully partitioned into a set of Q_T, which we denote by $\mathbb{Q}_{\mathbb{T}}$, such that subsets Q_T are pairwise disjoint. By assumption, the anonymity scheme (π, T, T^{-1}) achieves secure query $k-anonymity$ therefore $Q_T \subseteq s$ as asserted in Lemma 1. Since $s \subseteq \mathbb{Q}$, $Q_T \subseteq \mathbb{Q}$, $Q_T \subseteq s$, and the subsets Q_T are disjoint, then $s \setminus Q_T$ must be empty to have anonymity-sets $s \in \mathbb{S}$ disjoint. To prove this, assume otherwise, for the purpose of contradiction. That is, $s \setminus Q_T$ is not empty, and anonymity-sets $s \in \mathbb{S}$ are disjoint. Thus, given that $Q_{Ti} \subseteq s_i$ as in Lemma 1, there exists at least one destination node v_i that is in an anonymity-set s_i but its corresponding query q_i is not in Q_{Ti}. Since $\mathbb{Q}$ is fully partitioned into $\mathbb{Q}_\mathbb{T}$, then q_i must belong to another member of $\mathbb{Q}_\mathbb{T}$ say, $q_i \in Q_{Tj}$. But $Q_{Tj} \subseteq s_j$ by Lemma 1, where $s_j \neq\ s_i$. This implies that the corresponding destination node of q_i belongs to s_j. That is $v_i \in s_j$. Hence, $s_i \cap s_j = v_i$, thereby giving us the desired contradiction.

Consequently, the two conditions namely, the anonymity-sets are disjoint and $s = Q_T$, are necessary to preserve secure query $k - anonymity$ in the presence of eavesdropping adversary who is capable of multiple anonymity-sets interception.

□

Remark 2. *While the result $s = Q_T$ from Theorem 2, and Theorem 3 are similar, the distinction between their arguments is important. The former argues from totally an economical viewpoint, while the latter proves that the result is a prerequisite to assure secure query $k - anonymity$ against eavesdropping adversary with multiple anonymity-sets interception capability. A key element in Theorem 3, which is also of independent interest, is that it establishes the relationship between the set of bogus-destination queries S_b and the queries in Q_T, the input space of anonymity transformation T, specifically, $s = \{q, S_b\} = Q_T$.*

Corollary 2. *An anonymity scheme (π, T, T^{-1}), which achieves secure query $k-$anonymity against eavesdropping adversary with multiple anonymity-sets interception capability, provides a level-of-anonymity equal to the lowest size of the anonymity-sets, that is, $k = \min\limits_{s \in \mathbb{S}} |s|$, where $|s| = \Theta(k)$.*

Proof. The results follows intermediately from Theorem 3. Since $k = \min\limits_{Q_T \in \mathbb{Q}_\mathbb{T}} |Q_T|$ in secure query $k - anonymity$ as in definition 2, and $\min\limits_{Q_T \in \mathbb{Q}_\mathbb{T}} |Q_T| = \min\limits_{s \in \mathbb{S}} |s|$ by theorem 3 which asserts that $Q_T = s$, we have the desired result $k = \min\limits_{s \in \mathbb{S}} |s|$. Also, since $|Q_T| = \Theta(k)$ by definition 2, we have $|s| = |Q_T| = \Theta(k)$.

□

Based on the results obtained in Theorem 3 and Corollary 2 , the following conventions will be used alternatively when we consider secure query $k - \textit{anonymity}$:

- Since $k = |s|$, we will use the level-of-anonymity k and the size of anonymity-set $|s|$ alternatively as a measure to quantify the secure query $k - \textit{anonymity}$.
- Another implication of Theorem 3 is related to the partition algorithm π which constructs a set of anonymity-sets $\mathbb{S}$ out of the whole WSN. Since $s = Q_T$, from Theorem 3, each constructed anonymity-set $s \in \mathbb{S}$ is basically the input space of the anonymity transformation Q_T.

Chapter 4

Disjoint Anonymity-Sets (DAS): A Construction for a Secure K-Anonymous Query Scheme

In this chapter, the problem of designing a secure k-anonymous query scheme following Theorem 3 is formulated. Based on the generic construction presented in section 3.2, we propose an anonymity scheme, we call the Disjoint Anonymity-Sets (DAS), that achieves secure query k-anonymity as stated in Definition 2 in the presence of eavesdropping adversary armed with multiple anonymity-sets interception capability.

4.1 Disjoint Anonymity-Sets (DAS) Scheme

DAS is made up of one partition algorithm π and two anonymity transformations T and T^{-1}, as described below:

1. π is the partition algorithm that converts the whole WSN of size n into a union of disjoint non-empty anonymity-sets. Specifically, the inputs of the algorithm include WSN of size n, and $k \in [1, n]$ the desired level-of-anonymity. It outputs a partition of WSN containing a set disjoint anonymity-sets $\mathbb{S} = \{s_1, \ldots, s_{\lfloor n/k \rfloor}\}$, where $\cup_{s \in \mathbb{S}} = WSN$, and each $s \in S$ is of size $|s| = [k, 2k-1]$. The upper limit point of anonymity-set size, which is $2k-1$, captures that fact that n may not be divisible by k. Hence, there exists at least one anonymity-set of size in the closed interval $[k, 2k-1]$.

2. T is the anonymity transformation that maps every query q in its input space $Q_T \in \mathbb{Q}_{\mathbb{T}}$ to the corresponding anonymity-set $s \in \mathbb{S}$, i.e., $\forall q \in Q_T, \exists s = T(q)$, where $q \in s$. This means whenever the client wants to send a query q to a destination node, she is required to query each node in its corresponding anonymity-set s. Moreover $s = Q_T$, and thus the partition algorithm π defines not only the set of anonymity-sets $\mathbb{S}$ but also the set of input spaces of anonymity transformation $\mathbb{Q}_{\mathbb{T}}$.

   Since the DAS anonymity-sets are disjoint, the input spaces of T are disjoint too, which by nature resists all types of intersection and statistical attacks among different anonymity-sets [36, 37, 38]. This assures a certain level-of-anonymity regardless of the computational power of the eavesdropping adversary who is able to intercept the anonymity scheme for sufficiently long time.

3. T^{-1} is the inverse transformation that recovers the true-destination query from anonymity-set of the responses to the queries generated by anonymity transformation T. That is $T^{-1}s = T^{-1}(T(q)) = q$. Note that T and T^{-1} can both be performed by the client which meets the requirements of trust-none model (se section 1.2.1). It is interesting to note that the existence of T^{-1} verifies the correctness of the anonymity scheme in much the same manner that decryption does for the encryption scheme.

Theorem 4. *The DAS scheme is a secure k-anonymous query scheme with a level-of-anonymity equal to the lowest size of anonymity-sets, i.e.* $k = \min_{s \in \mathbb{S}} |s|$.

Proof. We must show that the DAS achieves level-of-anonymity $k = |Q_T|$ in the presence of eavesdropping adversary as stated in Definition 2. However this is immediately follows from the fact that DAS is designed to satisfy the conditions to achieve secure query k-anonymity namely, anonymity-sets are disjoint and each of them equal to the corresponding Q_T, as stated in Theorem 3. This is so because every $q \in Q_T$ is mapped to the same $s = Q_T$. Hence, it is indistinguishable in $k = |s| = |Q_T|$. Now, since n, WSN size, is not necessarily divisible by k, DAS partitioning leads to $|s| \in [k, 2k-1]$. We apply the well-known concept of weakest-link in security [83], the anonymity chain is only as strong as its weakest link to have $k = \min_{s \in \mathbb{S}} |s|$, which is also established in Corollary 2 .

□

4.2 Proposed DAS Scheme Design and Implementation

Detailed implementation to each of the sub-tasks under DAS in the context of square grid WSN is provided in this section.

4.2.1 The Partition Algorithm

We consider a source-routed square grid topology WSN ($\sqrt{n} \times \sqrt{n}$) that has been widely employed in the related research [8], where the position of each node is defined by the ordered pair of its Cartesian coordinates (i, j), where $i, j \in [1, \sqrt{n}]$. As in prior work [8], the client communicates directly with root-node $v_{1,1}$, which is the node at the upper left corner. The route between the root-node and any node is the shortest Manhattan distance. Note that Manhattan distance d_m between two nodes v_{i_1,j_1} and v_{i_2,j_2} is equal to the length of the path connecting them along the horizontal and vertical segments [88].

The following theorem formally defines the proposed partition algorithm, and its proof justifies its feasibility.

Theorem 5. *Given a $\sqrt{n} \times \sqrt{n}$ square grid connected undirected graph $G = (V, E)$, there exists a partition of V into disjoint sets of nodes $\mathbb{S} = \{s_1, s_2, \ldots, s_{\lfloor n/k \rfloor}\}$ each of size $\in [k, 2k-1]$, the DAS anonymity-sets, such that the following properties hold:*

1. *The set of nodes $\{v_1, v_2, \ldots, v_{|s_j|}\}$ in such anonymity-set s_j has the following total ordering. Given $i \in [1, |s_j| - 1]$, it holds that the Manhattan distance between nodes v_i and v_{i+1} is exactly 1.*

2. *v_1 (the first-node from the source along the route) is at the lowest vertical distance (y-coordinate) from the root-node $v_{1,1}$ compared to other nodes in such anonymity-set s_j.*

3. *Nodes $\{v_2, v_3, \ldots, v_{|s_j|}\}$ in such anonymity-set s_j is at a vertical distance of $\lceil |s|/\sqrt{n} \rceil$ from v_1.*

4. *The remainder out of partition n/k is distributed over anonymity-sets $\mathbb{S}$ using either First-Set-Spread (FSS), Equal-Spread (ES), or Random-Spread (RS) anonymity-sets construction methods (which will be defined right below).*

The properties (1) - (3) in the theorem ensure that the communication-cost is minimized when querying multiple sensor nodes in an anonymity-set, by which v_1 (the first-node along the route) is at the least vertical distance from $v_{1,1}$, while all the other nodes of the same anonymity-set is confined within a vertical distance of $\lceil |s|/\sqrt{n} \rceil$ from v_1.

The property (4) in the theorem ensures that the remainder nodes out of the partition n/k is distributed over the anonymity-sets by using one of the following three anonymity-sets construction methods. The first is called First-Set-Spread (FSS), which adds all the remaining nodes to the first set closet to the root-node in order to lower the communication-cost because the first set is the closet to the root-node. The second is called Equal-Spread (ES), which goes to the other extreme by distributing the remainder equally over all the anonymity-sets. Whereas the Random-Spread (RS) adopts the uniform random distribution to disseminate the remaining nodes randomly across anonymity-sets. The three methods will be examined in Chapter 6.

For the proof of existence of such disjoint anonymity-sets, our approach is to introduce a partition algorithm (Algorithm 1) that can be implemented in a square grid ($\sqrt{n} \times \sqrt{n}$) and its resultant partition meets all the properties listed in theorem 5. The algorithm calls four other procedures, namely $ConstructSpanningTree$ that constructs a comb-like spanning tree t of G, $ConstructHamiltonianPath$ that constructs the Hamiltonian path P of t, $FindAnonymitySetsSizes$ that determines the size of each anonymity-set, and $ConstructOneAnonymitySet$ that constructs one anonymity-set. A working example of the algorithm that uses First-Set-Spread (FSS) is shown in Fig. 4.1.

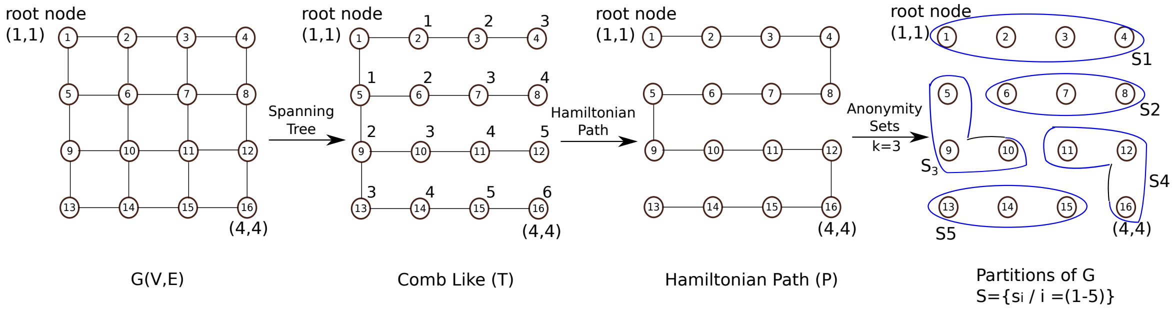


Figure 4.1: An example of running the partition algorithm shown in Algorithm 1 on 4×4 square grid connected undirected graph, and its procedures.We use Algorithm 1, with an input value of $k = 3$, to partition the graph $G = (V, E)$ on the far left into five disjoint anonymity-sets $\mathbb{S} = \{s_1, s_2, s_3, s_4, s_5\}$. First, We construct a comb-like spanning tree shown in the second graph from left using Breadth First Search (BFS) algorithm. We adopt a stepwise construction of the tree in which the sequence of steps is numbered on the diagram. Next we construct, in the third graph from left, Hamiltonian path in two iterations of $ConstructHamiltonianPath$ procedure. Finally, we use $ConstructOneAnonymitySet$ procedure, shown on the the right, to traverse the Hamiltonian path and generate the desired disjoint anonymity-sets. Since $n = 16$ is indivisible by $k = 3$ and we are using FSS anonymity-set construction method, the largest anonymity-set s_1 is the first to be created, and it is the closest to the root node $v_{1,1}$.

The Partition algorithm takes as inputs a $\sqrt{n} \times \sqrt{n}$ square grid connected undirected graph $G = (V, E)$, a desired level of anonymity $k \in [1, n]$, and an anonymity-sets construction method (ascm). In line 2, we call $ConstructSpanningTree$ to generate a comb-like spanning tree t of G. Then, we invoke $ConstructHamiltonianPath$ on t in line 3.

In $ConstructSpanningTree$, line 10, we invoke Breadth First Search (BFS) [89, 90] at root-node, $v_{1,1}$, which visits all of its adjacent nodes from left to right. Then for each of these adjacent nodes in turn, BFS visits their adjacent nodes from left to right if they are unvisited before, and continues in a similar manner. This forms a comb-like spanning tree as shown in Fig. 4.2 that has one spine of length $\sqrt{n}$ at $y = 1$ vertical line, and $\sqrt{n}$ teeth, each of length $\sqrt{n}$, along the x axis.

In $ConstructHamiltonianPath$, line 12, we convert the comb-like spanning tree t into a Hamiltonian path P. We accomplish this by removing some edges and replacing them with others in different locations. For doing this, In line 13, we iterate i through the range of rows in the square grid. That is, it takes on values from 1 through $\sqrt{n}$. In line 14, we limit i values to odd numbers. In the i^{th} iteration, we remove the edge between each two consecutive rows (i and $i + 1$) for odd values of i, and join them by an edge on the last column, as in lines 15 and 16.

To see why, refer to the comb-like tree t shown in the second graph from left of Fig. 4.1. To convert t into a Hamiltonian path, we need to: remove the edges joining the first and second rows on the first column, join them by inserting an edge on the last column, $\sqrt{n}^{th}$, leave the edge between the second and the third rows intact, and repeat the remove and insertion between the third and fourth rows. This is done in two iterations of the **for** loop in line 13, as indicated on the resultant Hamiltonian path in Fig. 4.1.

Back to the main body of Algorithm 1, we determine the size of each anonymity-set by

Algorithm 1 Partition Algorithm (π)

```
Input: $G = (V, E)$: $\sqrt{n} \times \sqrt{n}$ Square grid connected undirected graph , $k$: a desired level-of-anonymity, and $ascm$: an
    anonymity-sets construction method
Output: $\mathbb{S}$: the partition of $G$ into disjoint anonymity-sets
$n \leftarrow |V|$
$t \leftarrow ConstructSpanningTree(G)$
$P \leftarrow ConstructHamiltonianPath(t)$
$\{|s_i| \; |i \in [1, \lfloor n/k \rfloor]\} \leftarrow FindAnonymitySetsSizes(ascm, n, k)$
foreach $i \in [1, \lfloor n/k \rfloor]$ do
    $s_i \leftarrow ConstructOneAnonymitySet(P, |s_i|)$
    $P \leftarrow P$ with all nodes of $s_i$ removed
end for
return $\mathbb{S} \leftarrow \{s_i | i \in [1, \lfloor n/k \rfloor]\}$
procedure $ConstructSpanningTree(G')$
return $t' \leftarrow BFS(G')$ starting at root-node $v_{1,1}$ /* BFS is the Breadth First Search algorithm */
procedure $ConstructHamiltonianPath(t')$
/* $t'$ is a comb-like tree constructed by $ConstructSpanningTree$ procedure above */
for $i \leftarrow 1$ to $\sqrt{n}$ do
    if $i$ is odd then
        Remove the edge between nodes $(i, 1)$ and $(i + 1, 1)$ in $t'$
        Insert an edge between nodes $(i, \sqrt{n})$ and $(i + 1, \sqrt{n})$ in $t'$
    end if
end for
return $P' \leftarrow t'$
procedure $FindAnonymitySetsSizes(ascm', n', k')$
foreach $i \in [1, \lfloor n'/k' \rfloor]$ do
    $|s'_i| \leftarrow k'$
end for
$r \leftarrow n' \bmod k'$
if $ascm = FSS$ then
    $|s'_1| \leftarrow |s'_1| + r$
else if $ascm = ES$ then
    foreach $j \in [1, r]$ do
        $x \leftarrow j \bmod \lfloor n'/k' \rfloor$
        if $x = 0$ then
            $x \leftarrow \lfloor n'/k' \rfloor$
        end if
        $|s'_x| \leftarrow |s'_x| + 1$
    end for
else if $ascm = RS$ then
    foreach $j \in [1, r]$ do
        $x \leftarrow$ RandomNumberGenerator$(1, \lfloor n'/k' \rfloor)$
        $|s'_x| \leftarrow |s'_x| + 1$
    end for
end if
return $\{|s'_i| \; |i \in [1, \lfloor n'/k' \rfloor]\}$
procedure $ConstructOneAnonymitySet(P', l)$
if $l >$ number of nodes in $P'$ then
    return error
end if
Pre-order traverse $P'$ /* traverse from the the root-node, $v_{1,1}$ */
return a set containing the first $l$ nodes we encounter
```

calling *FindAnonymitySetsSizes* in line 4. This step is a prelude to assigning nodes to these anonymity-sets which is done in lines 5 to 8. In *FindAnonymitySetsSizes*, we set up an iterator, i, through all the anonymity-sets which takes values from 1 till $\lfloor n/k \rfloor$, line 21. In each iteration, we initialize the i^{th} anonymity-set, $|s'_i|$, to be equal to the desired level-of-anonymity, k'. This is the minimum size of an anonymity-set since the anonymity-set size is in $[k, 2k-1]$. In line 24, we calculate the remainder r out of the partition n/k using modulo operation.

Then, starting at line 25, we use **if-else if** statement to stress the fact that either one of the three anonymity-sets construction method (ascm), FSS, ES, or RS, is used. As mentioned in the proceeding paragraphs, these methods are used to distribute the remainder r over anonymity-sets. Simply, the FSS adds all the remaining nodes r to the first anonymity-set in line 26. Recall that ES attempts to add equal share of the remaining nodes r to each anonymity-set. Thus, we first set up an iterator, j, which takes values from 1 till the remainder r, line 28. In each iteration, we select an anonymity-set indexed by x sequentially, lines 29 to 32. Then, we increment its size by one node from the remaining nodes in line 33. Finally, for the RS, we again set up an iterator, j, which takes values from 1 till the remainder r, line 36. In each iteration, we increase the size of an anonymity-set whose index x is selected by a random number generator following the uniform distribution, lines 37 and 38.

In line 5 of the main body of the partition algorithm, we set up an iterator, i, through all the anonymity-sets which takes values from 1 till $\lfloor n/k \rfloor$. In each iteration, we call *ConstructOneAnonymitySet*, line 6, to construct the i^{th} anonymity-set. Then, we update the Hamiltonian path by removing the nodes of constructed i^{th} anonymity-set from it, as in line 7. Finally, after finishing the **for** loop in line 9, we return the partition $\mathbb{S}$ of G, which is a collection of anonymity-sets.

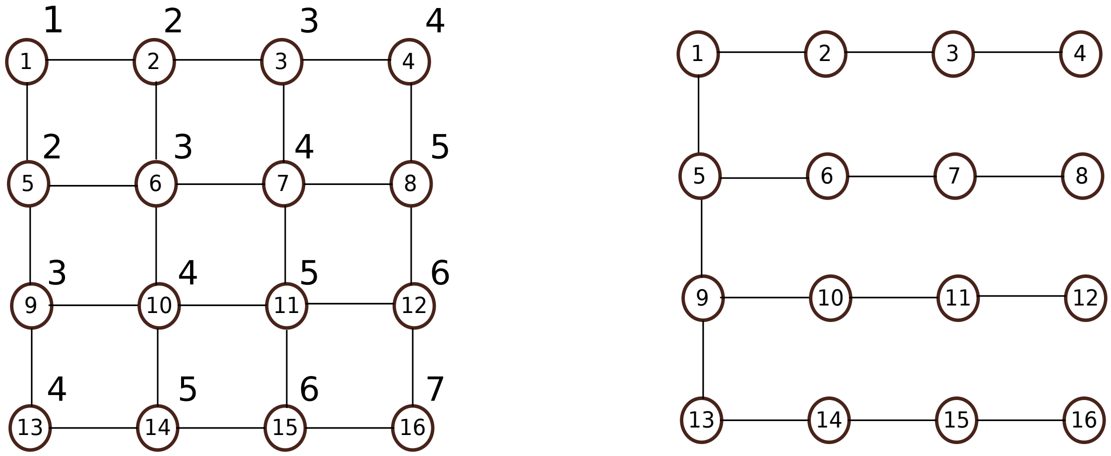
In $ConstructOneAnonymitySet$, we treat the Hamiltonian path p', which is one of the inputs in addition to an integer l, as a simple example of a tree. In line 46, we traverse the tree in a pre-order way [91]. That is, we recursively visit parent node from left to right before visiting their children node. In essence, this means we walk down the Hamiltonian path P' from its endpoint that is the root-node. In the last line, 47, we return the first l nodes we encounter to create one anonymity-set.

An important assertion is that the path from the root-node to any other vertex v in t is the shortest path from root-node to v in G. This can be easily verified via the fact that each connected graph has a spanning tree [90], as well as the fact that the shortest path spanning tree t can be built via Breadth First Search (BFS) [89, 90] which we include in the procedure $ConstructSpanningTree$ of the algorithm. Respectively, we derive the Hamiltonian path from the same spanning tree that we will use later for routing queries to the anonymity-sets, and collecting their responses, see section 4.2.2.

Note that the anonymity-sets are contiguous since all the nodes are possibly queried. Thus in the algorithm, we generate these anonymity-sets as a result of splitting a continuous path originating at the root-node to ensure the continuity requirement. On the other hand, since the path must be a Hamiltonian path that visits every node in WSN once, the disjointedness requirement is naturally met.

Time Complexity: We show that time of running Algorithm 1 on $G = (V, E)$ is linear in the number of nodes $|V| = n$. First, the Initialization in line 1 takes $O(1)$ time. $ConstructSpanningTree$ takes $O(|V| + |E|)$ with the BFS algorithm [91]. Now, in a $\sqrt{n} \times \sqrt{n}$ square grid graph, we observe that $|V| = n$, and $|E| = 2(n - \sqrt{n})$. Hence $ConstructSpanningTree$ takes $O(|V| + |E|) = O(n + 2(n - \sqrt{n})) = O(n)$.

In $ConstructHamiltonianPath$, the **for** loop costs $O(\sqrt{n})$ due to a constant time



(a) Stepwise construction of comb-like tress with BFS

(b) Comb-like Binary Spanning Tree

Figure 4.2: An example of comb-like rooted binary spanning tree construction algorithm using BFS for 4×4 square grid connected undirected graph. (a) Stepwise construction of the tree in which the sequence of steps is numbered on the diagram, (b) the resultant tree

consumed by the execution of the body of the loop. *FindAnonymitySetsSizes* in line 4 yields $O(n)$ of running time. This is so because its cost is mainly of two parts namely, the **foreach** loop and the **if-else if** statement, and each one of them is in $O(n)$ as follows. The **foreach** in line 21 costs $\lfloor n/k \rfloor$ since its body takes constant time. But $k \in [1, n]$, therefore **foreach** is in $O(n)$. The FSS case, line 25, of **if-else if** costs a constant time, i.e $O(1)$. The ES case costs the same of its **foreach** loop in line 28 which performs a constant time block r times. Recall that r, the remainder out of the devision n/k, is in $O(k)$ which is in turn in $O(n)$. Hence the ES case is in $O(n)$. Finally, the RS case is in $O(n)$ because its **foreach** performs a constant time block of code r times. The **foreach** loop in lines 5–8 performs *ConstructOneAnonymitySet* and the remove of s_i nodes operation, line 7, $\lfloor n/k \rfloor$ times. Considering *ConstructOneAnonymitySet*, lines 42–45 costs a constant time, and the Pre-order, line 46, traverse only $|s_i|$ nodes of the

Hamiltonian path. The remove operation, line 7, removes $|s_i|$ nodes from the Hamiltonian path. Since $|s_i| \in [k, 2k-1]$ (see section 4.1), each of *ConstructOneAnonymitySet* and the remove operation, line 7, takes time $\Theta(k)$. Therefore, we can state the running time of the **for** loop in lines 5–8 as $O(\lfloor n/k \rfloor \times (k+k) = O(n/k \times 2k) = O(n)$. Consequently, the total running time of the Partition algorithm, which is equal to the sum of the time taken by *ConstructSpanningTree*, *ConstructHamiltonianPath*, *FindAnonymitySetsSizes*, and the **foreach** loop in lines 5–8, is $O(n + \sqrt{n} + n + n) = O(n)$.

Now we are able to prove the existence of a valid partition according to Theorem 5, which also proves the correctness of the Partition algorithm.

Proof for Theorem 5. Firstly we argue that by implementing the Partition Algorithm 1, the generated anonymity-sets are disjoint since each of them is obtained by splitting a Hamiltonian path of length n into path segments of size in $[k, 2k-1]$. Then we argue that the four properties in theorem 5 are satisfied as follows. Property (1) is satisfied due to the fact that the Manhattan distance between nodes v_i and v_{i+1} is exactly 1, for $i \in [1, |s_j| - 1]$, since they are successive nodes on the Hamiltonian path. To prove the feasibility of the second property, we specify v_1 of each anonymity-set to be the endpoint of the corresponding path segment that is closest to the root-node along the Hamiltonian path. The property is satisfied because the root-node, $v_{1,1}$, is at $y = 1$, the lowest y-coordinate. For example, v_1 becomes the root-node itself in the anonymity-set that contains the root-node, wherein the y-coordinate of v_1 is the lowest among other nodes in each anonymity-set, and nodes $\{v_2, v_3, \ldots, v_{|s_j|}\}$ are located on the same row of v_1 or below it. Since the length of each row is $\sqrt{n}$, the number of rows that the anonymity-set spans is in $\Theta(\lceil |s|/\sqrt{n} \rceil)$, which satisfies the third property. The fourth property is naturally satisfied because in Algorithm 1, the remainder nodes are distributed over the anonymity-sets using one of the anonymity-sets construction methods namely, FSS, ES, or RS.

□

4.2.2 Query Routing

To query a desired (by the client) destination node v_d in WSN anonymously using DAS scheme, we need to route a query to every member of the anonymity-set s_j to which v_d belongs. The proposed routing algorithm achieves this by building a routing substrate based on the comb-like shortest-path spanning tree provided by the partitioning Algorithm 1. By launching a single query that visits all nodes of s_j to which v_d belongs, the query responses of all the nodes are collected in a piggyback manner [9], where each node in the anonymity-set attaches its response data with its ID to the query packet. This is possible mainly because the nodes in DAS anonymity-sets are one hop from each other.

Besides the WSN graph G, and the queried node v_d, one of the inputs to our Source-Route Construction algorithm, listed in Algorithm 2, is the partition of G obtained using the Partition algorithm, listed in Algorithm 1. As mentioned in section 4.2.1, the partition consists of a family of disjoint anonymity-sets $\mathbb{S} = \{s_i | i \in \lfloor n/k \rfloor\}$. Each anonymity-set $s_i \in \mathbb{S}$ has its own first-node v_1, and last-node $v_{|s_i|}$.

We present an example of running the routing algorithm in Fig. 4.3. The proposed Source-Route construction, namely Algorithm 2, is given as follows.

The output of Algorithm 2 is the list h_j which contains the source-route header information required by the DAS anonymity scheme. Each entry in h_j consists of two elements: the node ID value denoted by ID_v, and the node action value denoted by $Action_v$. When a query request is received, we first invoke *FindAnonymitySet*, in line 1, to find the the index $j \in [1, \lfloor n/k \rfloor]$ of the anonymity-set s_j to which v_d belongs. In *FindAnonymitySet*, line 10,

Algorithm 2 Source-Route Construction

Input: $G = (V, E)$: $\sqrt{n} \times \sqrt{n}$ connected undirected graph, $v_d \in V$: the queried destination node, and $\mathbb{S} = \{s_i | i \in [1, \lfloor n/k \rfloor]\}$: the anonymity-sets obtained from running Algorithm 1

Output: h_j: the source-route header information which is an ordered list of node IDs with their action values, and s_j: the corresponding anonymity-set of v_d, where $j \in [1, \lfloor n/k \rfloor]$

```
$j \leftarrow FindAnonymitySet(v_d)$
$(x_1, y_1) \leftarrow$ coordinates of $v_1$ of $s_j$ /* The coordinates of first-node of the anonymity-set $s_j$ */
$(x_{|s_j|}, y_{|s_j|}) \leftarrow$ coordinates of $v_{|s_j|}$ of $s_j$ /* The coordinates of last-node of the anonymity-set */
Initialize $h_j$ to an empty list
$t \leftarrow ConstructSpanningTree(G)$
$SourceRoutePart1(t, h_j, x_1, y_1)$
$SourceRoutePart2(s_j, h_j)$
$SourceRoutePart3(t, h_j, x_{|s_j|}, y_{|s_j|})$
return $h_j, s_j$
procedure $FindAnonymitySet(v')$
for $i \leftarrow 1$ to $\lfloor n/k \rfloor$ do
    if $v' \in s_i$ then
        $j' = i$, break
    end if
end for
return $j'$
procedure $ConstructSpanningTree(G')$
return $t' \leftarrow BFS(G')$ starting at root-node $v_{1,1}$ /* BFS is the Breadth First Search algorithm */
procedure $SourceRoutePart1(t', h'_j, x'_1, y'_1)$
for $i \leftarrow 1$ to $y'_1 - 1$ do
    $AppendForwardingNode((1, i), h'_j)$
end for
for $i \leftarrow 1$ to $(x'_1 - 1)$ do
    $AppendForwardingNode((i, y'_1), h'_j)$
end for
procedure $SourceRoutePart2(s'_j, h'_j)$
foreach node $v \in s'_j$ do
    $ID_v \leftarrow$ ID of node $v$
    $Action_v \leftarrow 1$ /* piggyback action */
    Append $\langle ID_v, Action_v \rangle$ to $h'_j$
end for
procedure $SourceRoutePart3(t', h'_j, x'_{|s_j|}, y'_{|s_j|})$
for $i \leftarrow x'_{|s_j|} - 1$ to 1 do
    $AppendForwardingNode((i, y'_{|s_j|}), h'_j)$
end for
for $i \leftarrow y'_{|s_j|} - 1$ to 1 do
    $AppendForwardingNode((1, i), h'_j)$
end for
procedure $AppendForwardingNode((x, y), h''_j)$
$ID_v \leftarrow$ ID of the node corresponds to point $(x, y)$ in the plane
$Action_v \leftarrow 0$ /* forward action */
Append $\langle ID_v, Action_v \rangle$ to $h''_j$
```

an iterator i is set up through all the anonymity-sets to check which one contains the queried node, v_d. In lines, 2 and 3, we specify x and y coordinates of the first-node as $v_1(x_1, y_1)$, and the last-node as $v_{|s_j|}(x_{|s_j|}, y_{|s_j|})$, of the discovered anonymity-set s_j. In line 4, we initialize h_j, which contains the source-route header information, to an empty list. In line 5, we call the procedure $ConstructSpanningTree$ on G to generate a comb-like shortest-path spanning tree t, and then invoke the following three procedures: $SourceRoutePart1$, $SourceRoutePart2$, $SourceRoutePart3$ in lines 6, 7, and 8 respectively, on t, and s_j as input arguments.

$SourceRoutePart1$ constructs the route to $v_1(x_1, y_1)$, which is the first-node of s_j, by traversing the comb-like shortest-path spanning tree t using two **for** loops. In the first loop, line 20–22, we traverse the tree vertically by increasing the y coordinate along the spine of the comb, i.e. $x = 1$ vertical line, starting from root-node $v_{1,1}$, until reaching the tooth of the comb with a y coordinate equal to $(y_1 - 1)$. In the second **for** loop, line 23–25, we traverse the tree along the next tooth, i.e. the one with a y coordinate equal to y_1, horizontally by increasing the x coordinate till we hit v_1. Each time through any of the two **for** loops, we append a forwarding node entry to h_j by invoking another procedure, namely $AppendForwardingNode$ in line 21 and 24, which specifies the two elements of the appended node entry in h_j as follows: ID of the node, and action value of 0, which denotes a forward action only. Then we append the forwarding node entry.

In $SourceRoutePart2$ (line 26–31), we iterate through the nodes of the anonymity-set s_j. In each iteration of the **foreah** loop, we specify the two element of the appended node entry as follows: the ID of the node, and the action value of 1, which denotes a respond action in a piggyback manner. Then we append the node to h_j.

In $SourceRoutePart3$, we construct the route from the last node of s_j namely, $v_{|s_j|}$ to the root-node $v_{1,1}$. To achieve this, we traverse t using two **for** loops. In the first loop, line

33–35, we traverse t horizontally by decreasing the x coordinate, starting from $x = x_{|s_j|} - 1$ till we reach the spine at $x = 1$. Then, in the second **for** loop on line 36–38, we traverse t vertically by decreasing the y coordinates along the spine, starting from $y_{|s_j|} - 1$ till we reach $v_{1,1}$. Similar to $SourceRoutePart1$, in each iteration of any of the **for** loops, we invoke $AppendForwardingNode$ to append a forwarding nodes entry to h_j after specifying its two element: ID of the node, and action value of 0, which denotes a forward action only.

Time Complexity: The Source-Route Construction algorithm, listed in Algorithm 2, runs in time linear with the size of the square grid WSN, $|V| = n$. To see that, firstly we observe that $FindAnonymitySet$ takes at most $O(\lfloor n/k \rfloor) = O(n)$. The assignments and initialization, lines 2, 3, and 4, takes constant time, $\Theta(1)$. As we establish in section 4.2.1, the running time of $ConstructSpanningTree$ on a square grid graph is in $O(|V| + E||) = O(n)$. Each of the two **for** loops of $SourceRoutePart1$ takes $O(\sqrt{n})$. This is so because the first loop iterates along the spine of the comb-like spanning tree, and the second iterates along one of its teeth, and both are in $O(\sqrt{n})$ as mentioned in section 4.2.1.

The same argument is valid for the running time of $SourceRoutePart3$ with the parts of the comb-like tree along which the two **for** loops iterate are reversed. Therefore, $SourceRoutePart3$ takes $O(\sqrt{n})$. Lastly, $SourceRoutePart2$ runs in time linear with the size of the anonymity-set of the queried node v_d, i.e. it is in $\Theta(|s_j|)$. Since $|s_j| \in [k, 2k-1]$, and $k \in [1, n]$, thus $SourceRoutePart3$ takes time $O(n)$ in the worst case. Therefore, the total running time of the Source-Route Construction algorithm, listed in Algorithm 2, which is equal the sum of time taken by $FindAnonymitySet$, $ConstructSpanningTree$, $SourceRoutePart1$, $SourceRoutePart2$,and $SourceRoutePart3$ is $O(n + n + \sqrt{n} + n + \sqrt{n}) = O(n)$.

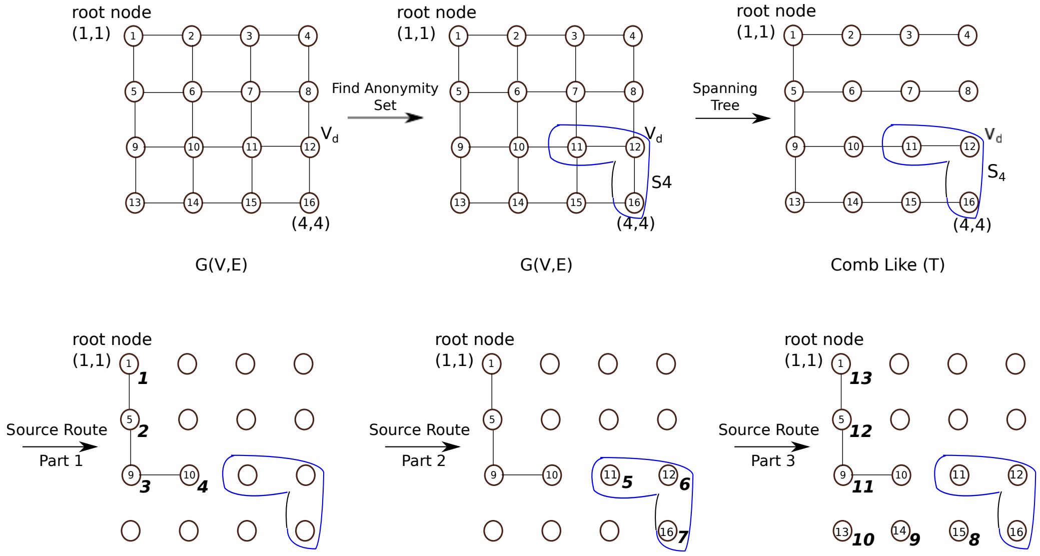


Figure 4.3: An example of running the Source-Route Construction algorithm shown in Algorithm 2 on a 4 × 4 square grid connected undirected graph shown on the far left, and its procedures: *FindAnonymitySet*, *ConstructSpanningTree*, *SourceRoutePart*1, *SourceRoutePart*2, and *SourceRoutePart*3. We use Algorithm 2 with an input value of $v_d = 12$, i.e. the ID of the queried node is 12, to construct the source-route required to query the whole anonymity-set $s_4 = \{11, 12, 16\}$ to which v_d belongs. The other input to Algorithm 2 is the family of anonymity-sets that results from Algorithm 1, as shown on the right of Fig. 4.1. First, we determine s_4 as the anonymity-set of $v_d = 12$ (the second graph from left) using *FindAnonymitySet* procedure. For s_4, $v_1 = 11$, and $v_{|s_j|} = 16$. Next, we construct, as in the third graph from left, a comb-like shortest path spanning tree using BFS algorithm. Finally, we use *SourceRoutePart*1, *SourceRoutePart*2, and *SourceRoutePart*3 to add node IDs to the source route header information h_4 along with their action values in the following order: $\langle 1, 0\rangle$, $\langle 5, 0\rangle$, $\langle 9, 0\rangle$, $\langle 10, 0\rangle$, $\langle 11, 1\rangle$, $\langle 12, 1\rangle$, $\langle 16, 1\rangle$, $\langle 15, 0\rangle$, $\langle 14, 0\rangle$, $\langle 13, 0\rangle$, $\langle 9, 0\rangle$, $\langle 5, 0\rangle$, $\langle 1, 0\rangle$. Action value of 0 implies forward only, and 1 implies piggyback response data then forward. We adopt a stepwise construction of the source-route in which the sequence of the steps of adding nodes to h_4 is marked on Fig. 4.3 .

4.2.3 The Querying Protocol

Our querying protocol, which implements the DAS anonymity scheme (π, T, T^{-1}), defines the rules for sending the client's query to a desired destination node v_d, and collecting its response anonymously. The protocol involves the following parties: the client $\mathcal{C}$, the sensor-cloud $\mathcal{S}$, and the WSN nodes V. $\mathcal{C}$ is the initiator of the protocol since she is the party who decides when to send a new query, the event that starts the execution of the protocol. Based on their action values in the header of the query packet, WSN nodes either forward the query packet to the next hop in the source-route header information, or piggyback their response data before forwarding.

In Protocol 1, we describe, in detail, how the parties engage in the querying protocol by exchanging messages, and acting upon receiving them. An instance of executing the querying protocol is depicted in Fig. 4.4. In the following, we identify some relevant aspects of the protocol that comply with the requirements of our trust model as specified in section 1.2.1.

Recall from section 1.2.1, in our trust-none model, the owner of the sensitive information namely, the client $\mathcal{C}$, trusts no other player in the anonymity scheme. In other words, the model allows the adversary to be outsider or any one of the querying protocol players except the owner of the sensitive information $\mathcal{C}$. Moreover, the scheme remains secure against the collusion of any of its outsider or insider adversaries. Consequently, in addition to the partition algorithm π, all of the anonymity transformations (T, and T^{-1}) are performed exclusively at the client $\mathcal{C}$. This is illustrated at the beginning of Protocol 1 where Algorithm 2, which utilizes the Partition algorithm π, is used to find the anonymity-set s_j corresponds to the queried node v_d, and construct the source-rote header information h_j. In essence, this means that the anonymity transformation T and the partition algorithm

π are run entirely at $\mathcal{C}$ before even contacting the sensor-cloud $\mathcal{S}$. Further, at the end of Protocol 1, the inverse anonymity transformation T^{-1} is performed at $\mathcal{C}$.

Protocol 1 The Querying Protocol

```
upon receiving a request from the client $\mathcal{C}$ to query a destination node $v_d$:
 Using Algorithm 2:
   Find anonymity-set $s_j$ of $v_d$
   Generate corresponding source-route header information $h_j$
   Initialize $R$ to an empty list of size $|s_j|$
/*  $R$ stores piggybacked response data */
 Create a query packet $\rho$:
   Insert the client query data, and $R$ into the payload of $\rho$
   Insert $h_j$ into the header of $\rho$
 Send $\rho$ to the Sensor-Cloud $\mathcal{S}$
upon receiving $\rho$ by $\mathcal{S}$ from $\mathcal{C}$:
 Forward $\rho$ to WSN through $v_{1,1}$ (the root-node of WSN)
upon receiving $\rho$ by an arbitrary WSN node $v$:
if $v$ is not the first node entry in $h_j$ then
   Drop $\rho$
else
   Remove $v$'s entry $\langle ID_v, Action_v \rangle$ from $h_j$
   if $Action_v = 0$ then
      $\rho \leftarrow \rho$ with $v$'s entry removed from $h_j$
      $Forward(\rho)$
   else if $Action_v = 1$ then
      Append $v$'s response data $d_i, i \in [1, n]$ to $R$ in $\rho$
      $\rho \leftarrow \rho$ with $v$'s entry removed from $h_j$
      $Forward(\rho)$
   end if
end if
upon receiving $\rho$ by $\mathcal{S}$ from $v_{1,1}$:
 Send back to $\mathcal{C}$
upon receiving $\rho$ by $\mathcal{C}$ from $\mathcal{S}$:
 Extract $v_d$'s response data from $R$
 Drop the rest of $\rho$
procedure $Forward(\rho')$
if $h_j$ is empty then
   Send back to $\mathcal{S}$
else
   Forward $\rho'$ to the next node in $h_j$
end if
```

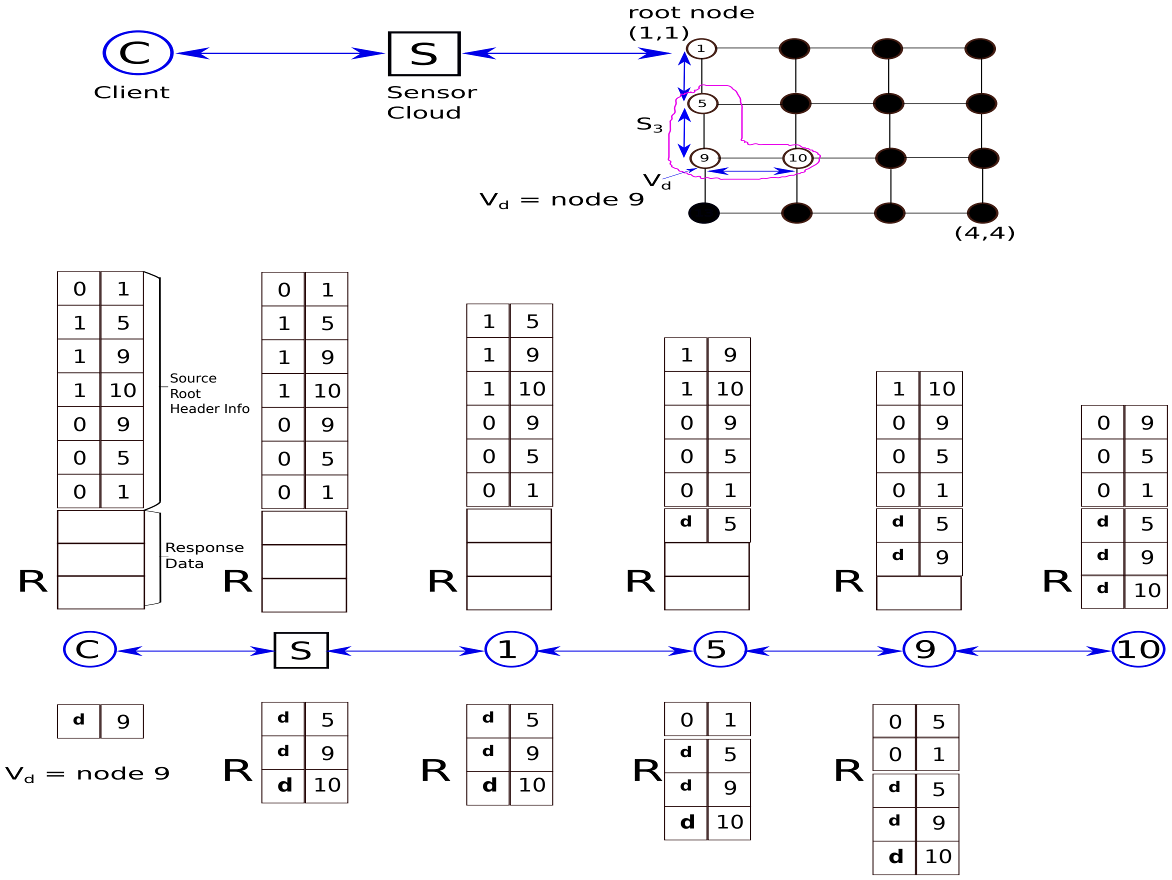


Figure 4.4: An example of executing the querying protocol shown in Protocol 1 using a 4×4 square grid connected undirected graph. The action field is 1 for the anonymity-set members so they attach their response data R to the packet in a piggy-back manner. First, the protocol uses Algorithm 2 to find the anonymity-set $s_3 = \{5, 9, 10\}$ of the queried node $v_d = 9$, as shown on the right of Fig. 4.1. Using the same algorithm, it creates the required source-route header information $h_3 = \langle 1, 0\rangle, \langle 5, 1\rangle, \langle 9, 1\rangle, \langle 10, 1\rangle, \langle 9, 0\rangle, \langle 5, 0\rangle, \langle 1, 0\rangle$. Then, it inserts h_3 along with the piggyback response data structure R, which is of size equal to $|s_3| = 3$, and the query data into the query packet. The upper part of the figure shows the path taken by the query packet, which starts and ends at the client, in order to collect responses from all nodes in s_3 as required by the DAS anonymity scheme. The lower part of the figure tracks down the changes in the query packet header, specifically in h_3, and the piggyback response data collection. At the end, the client keeps d_9, the response data of the queried node $v_d = 9$, and drops the rest of the packet.

Chapter 5

Cost Analysis of Secure K-Anonymous Query Schemes

It is important to know how much communication-cost is needed, on average and in the worst-case, to guarantee that the desired level-of-anonymity k is achieved using a secure k-anonymous query scheme such as DAS (see Chapter 4). In this chapter, we analyze the trade-off between incurred communication-cost and offered query-anonymity in a square grid WSN that is the main issue in communication-based class of anonymous communication schemes (see section 2.2).

Since DAS facilitates the measurement of query-anonymity by having the level-of-anonymity equal to the minimum size of anonymity-sets, i.e. $k = \min_{s \in \mathbb{S}} |s|$, which is established in Theorem 4, we will be presenting, in the rest of this chapter, a model to measure the incurred communication-cost (section 5.1), and an analysis of the cost-anonymity trade-off (section5.2).

5.1 Communication-Cost Measurement

Assume the query and response data are of a constant and uniform size, and the cost of sending this constant amount of data over one hop is constant as well, i.e. $\Theta(1)$. We consider a square grid homogeneous WSN of size n nodes. The position of each node is defined by the ordered pair of its Cartesian coordinates (i, j), where $i, j \in [1, \sqrt{n}]$. In our setting, the client communicates directly with root node at $(1, 1)$, namely $v_{1,1}$. A route of the shortest Manhattan distance from the root node $v_{1,1}$ to each destination node is predefined. Manhattan distance between two nodes is equal to the sum of the absolute differences of their Cartesian coordinates [88]. By modeling a square grid WSN as a connected undirected graph, the Manhattan distance corresponding to the longest shortest path between any pair of nodes is equal to $2(\sqrt{n}-1)$. It is called the diameter of the graph, which we denote by d. The lemmas below facilitate the calculation of communication-cost along the three parts of the routing path which are discussed in section 4.2.2.

Lemma 2. *The communication-cost* (c) *of transmitting a packet of a constant size data along a source-routing path of length* (l) *in* $G = (V, E)$ *is asymptotically quadratic with respect to the path's length. That is,* $c = \Theta(l^2)$.

Proof. Assume source routing is implemented where the path for each packet is kept in the packet header. Obviously, the initial length of the source-route entries record (number of nodes) in the packet header is $l + 1$, where l is the path length from source to destination node. We add 1 to count for the source node. Since the path length shrinks to one entry at the final destination, we can asymptotically abstract the incurred communication-cost from transmitting such a packet as: $c = \Theta(l+1) + \Theta(l) + \cdots + \Theta(1) = \Theta(l) + \cdots + \Theta(l) = \Theta(l^2)$.

□

Lemma 3. *Assume the response data of each node in the anonymity-set is of a constant size* $\Theta(1)$*, the communication-cost c of collecting responses from all nodes in the anonymity-set of a secure k-anonymous query scheme is asymptotically quadratic with respect to the offered level-of-anonymity k. That is,* $c = \Theta(k^2)$*.*

Proof. With the fact that transmission over one hop incurs a constant communication-cost $\Theta(1)$, the communication-cost is incremented by $\Theta(1)$ over every next hop in the anonymity-set. Hence, the total communication-cost of the piggyback is:$\Theta(|s|) + \Theta(|s| - 1) + \cdots + \Theta(1) = \Theta(k) + \Theta(k-1) + \cdots + \Theta(1) = \Theta(k) + \Theta(k) + \cdots + \Theta(k) = \Theta(k^2)$. The last equality uses the results we have established in Corollary 2 namely, $|s| = \Theta(k)$ in the secure k-anonymous query scheme.

□

5.2 Average and Worst-Case Analysis

In this section, we present our asymptotic analysis for the trade-off between query-anonymity and incurred communication-cost in the source-routed square grid WSN. We show that the diameter of the graph, d, which is the longest shortest path between any pair of nodes, plays the role of an inflection point. At this inflection point, the necessary and sufficient communication-cost changes from asymptotic dependence on d^2, to quadratic asymptotic dependence on the offered level-of-anonymity, k^2. The basic intuition behind that is the communication-cost can be expressed asymptotically in terms of the sum of d^2 and k^2. Thus, when k is much less than d, i.e. $k \in [1, \Theta(d)]$, the cost grows asymptotically with d^2, and vice versa.

We present our theoretical results for the average and worst-case cost analysis of the trade-off in the following asymptotic assertions.

5.2.1 Average-Case Analysis

Theorem 6. *To achieve secure query $k - anonymity$ of $k \in [1, \Theta(d)]$ in a source-routed square grid WSN of size n whose diameter is d, an average-case cost of $c_a = O(d^2)$ is sufficient, and $c_a = \Omega(d^2)$ is necessary.*

Proof. Sufficiency: Recall that the proposed construction, which is presented in Chapter 4, is a provably secure k-anonymous query scheme by Theorem 3. Therefore, In order to prove this part, it is enough to show that an average cost of $c_a = O(d^2)$ is sufficient for it to achieve a level-of-anonymity of $k \in [1, \Theta(d)]$. Let V_d be the random variable whose value equals the destination node v_d that is queried by the client. In fact, we have no idea about how the client selects v_d, nor do we have any control over the selection process. Consequently, it is natural to assume that all WSN nodes $v_1, \ldots, v_n$ are equally likely to be a destination node v_d of the client query. Thus, we have $Pr[V_d = v_{di}] = 1/n$, where $v_{di} \in V$ is an arbitrary node in WSN, and $i \in [1, n]$. To this end, we calculate the average-case cost c_a by applying the definition of the expected value [92], $c_a = \sum_{i=1}^{n} c_i \; Pr[V_d = v_{di}]$ Where c_i denotes the total communication-cost of querying an arbitrary node v_{di}, and collecting its response anonymously. It comprises both the query and response routing cost and piggyback data collection cost.

It remains to prove that $c_i = O(d^2)$. First, we consider that the client seeks to query an arbitrary node of WSN v_{di} which belongs to an anonymity-set s_j anonymously. To bound the incurred communication-cost from above, we examine the path from the client to the anonymity-set s_j via the sensor-cloud. As specified in section 4.2.2, our routing path to v_{di}

is split into three parts. We claim that the total length of the three parts of the constructed route is in $O(d)$. It is so because each of the first and third parts are the shortest paths between two nodes namely, $v_{1,1}$,v_1, and $v_{1,1}$, $v_{|s_j|}$ correspondingly, where v_1, and $v_{|s_j|}$ are the first-node and the last-node of s_j respectively. Nonetheless, the diameter of the graph, d, is the longest shortest path between any pair of nodes. Thus, the first and third parts are in $O(d)$. For the second part, its path length is equal to the size of the anonymity-set, $|s_j|$, $|s_j|$ is in $\Theta(k)$ by Corollary 2, and $k \in [1, \Theta(d)]$ by assumption. Thus, the total length of the query and response routing path is in $O(d)$. By Lemma 2, we conclude that the incurred routing cost from the three parts together is in $O(d^2)$. Similarly, the piggyback data collection cost is in $O(d^2)$, since it is in $\Theta(k^2)$ by Lemma 3, and $k \in [1, \Theta(d)]$ by assumption. At last, the collected data of size $|s_j| = \Theta(k)$ is forwarded along the third part of the route, which is of length $O(d)$, back to the client. This incurs a worst-case cost of $O(k\ d)$, but since $k \in [1, \Theta(d)]$, the incurred worst-case cost is in $O(d^2)$. Consequently, total communication-cost which is the sum of routing and data collection cost is in $O(d^2)$.

Now we can compute the average-case cost as, $c_a = \sum_{i=1}^{n} c_i\ Pr[V_d = v_{di}] = n\ O(d^2)\ 1/n = O(d^2)$.

Necessity: we prove by contradiction. Assume otherwise that there exists a secure k-anonymous query scheme that achieves a level-of-anonymity $k \in [1, \Theta(d)]$, and requires a communication-cost of $o(d^2)$ in the average-case (note that $o(d^2)$ is the complement of $\Omega(d^2)$). To bound the necessary average cost of such a scheme from below, we compute the cost of the least amount of data, 1 bit, that is sent to each node of WSN and back. This refers to the case where no anonymity is offered, i.e. $k = 1$. The total cost therefor is merely the routing cost. Then we average across all nodes of WSN. As in the proof of the sufficiency part, we apply the definition of the expected value [92] namely, $c_a = \sum_{i=1}^{n} c_i\ Pr[V_d = v_{di}]$, where $Pr[V_d = v_{di}] = 1/n$ as discussed in the sufficiency part of this theorem, and c_i denotes the communication-cost of sending 1 bit to node v_{di}. We argue that $c_i = \Omega(d^2)$. The reason is that the Manhattan distance values from root-node to WSN nodes are $d, d-1, d-2, \ldots,$ and so on. By choosing suitable positive constant values, it is easy to show that these distances are in $\Omega(d)$. Thus, By Lemma 2, we conclude that the incurred communication-cost is in $\Omega(d^2)$. Plugging in theses values into the expected value definition, the result follows. That is, $c_a = \sum_{i=1}^{n} c_i \; Pr[V_d = v_{di}] = n \; \Omega(d^2) \; 1/n = \Omega(d^2)$, which gives us the desired contradiction.

□

Theorem 7. *To achieve secure query $k-anonymity$ of $k \in [\Theta(d), n]$ in a source-routed square grid WSN of size n whose diameter is d, an average-case cost of $c_a = O(k^2)$ is sufficient, and $c_a = \Omega(k^2)$ is necessary.*

Proof. Sufficiency: We carry out the proof of this part similar to the sufficiency proof of Theorem 6. As an example of secure k-anonymous query schemes, we analyze the cost required by the proposed construction (see Chapter 4) to achieve a level-of-anonymity of $k \in [\Theta(d), n]$. To compute the sufficient average cost, we use the definition of the expected value [92], $c_a = \sum_{i=1}^{n} c_i \; Pr[V_d = v_{di}]$, where, $Pr[V_d = v_{di}] = 1/n$ as discussed in Theorem 6, and c_i denotes the total communication-cost of querying an arbitrary node v_{di}, and collecting its response anonymously. It incorporates both the routing cost and piggyback data collection cost. To find c_i, we pursue the same logic of the corresponding sufficiency proof of Theorem 6. Thus, we consider the destination node of the client anonymous query to be an arbitrary node of WSN, v_{di}. Recall that our routing path is split into three parts, as in section 4.2.2. The path length of the first and third parts of the route is in $O(k)$ since each part is in $O(d)$, and $k \in [\Theta(d), n]$ by assumption. Therefore, the length of the three parts of the route is in $O(k)$ since the length of the second part is equal to $|s_j|$ which is in $\Theta(k)$ from Corollary 2. By Lemma 2, the routing cost is in $O(k^2)$. The data collection is in $O(k^2)$ as we establish in Lemma 3. Additionally, the collected data of size $O(k)$ incurs a cost of $O(k^2)$ when it travels back to the client via sensor-cloud along the third part of the route, which is of length $O(d)$. This is because $k \in [\Theta(d), n]$ by assumption, and the incurred cost is in $O(k\ d)$. Consequently, the total communication-cost of an arbitrary node in WSN $c_i \in O(k^2)$. Now we are able to compute the average cost to be, $c_a = \sum_{i=1}^{n} c_i\ Pr[V_d = v_{di}] = n\ O(k^2)\ 1/n = O(k^2)$.

Necessity: For the purpose of contradiction, suppose there exists a secure k-anonymous query scheme that achieves a level-of-anonymity $k \in [\Theta(d), n]$, and requires a communication-cost of $o(k^2)$ in the average-case. To analyze its necessary cost in the average-case, we adopt the definition of the expected value [92], $c_a = \sum_{i=1}^{n} c_i\ Pr[V_d = v_{di}]$, where, $Pr[V_d = v_{di}] = 1/n$ as discussed in Theorem 6, and c_i denotes the necessary communication-cost required by this hypothesized scheme to query an arbitrary node $v_{di} \in V$, and collecting its response anonymously. To achieve this, the scheme must be able to query all nodes of the anonymity-set s_j to which v_{di} belongs. It is clear that c_i is bounded from below by the cost of sending a minimum of 1 bit to all $|s_j|$ nodes of the anonymity-set to which v_{di} belongs. Hence, The source route length must be at least equal to $|s_j|$. By Lemma 2, The cost of this communications is in $\Omega(|s_j|^2)$. Therefore, the necessary communication-cost c_i is in $\Omega(k^2)$ using the result from corollary 2. Now we are able to compute the necessary average cost to be, $c_a = \sum_{i=1}^{n} c_i\ Pr[V_d = v_{di}] = n\ \Omega(k^2)\ 1/n = \Omega(k^2)$. This gives us the desired contradiction.

□

5.2.2 Worst-Case Analysis

Theorem 8. *To achieve secure query $k - anonymity$ of $k \in [1, \Theta(d)]$ in a source-routed square grid WSN of size n whose diameter is d, a worst-case cost of $c_w = O(d^2)$ is sufficient, and $c_w = \Omega(d^2)$ is necessary.*

Proof. Sufficiency: recall that DAS is a provably secure k-anonymous query scheme, as we establish in Theorem 4. Thus, to prove the sufficiency, it is enough to show that a worst-case communication-cost of $O(d^2)$ is sufficient for DAS to achieve a level-of-anonymity of $k \in [1, \Theta(d)]$. First, we observe that the worst-case cost is incurred when the client seeks to query the farthest node from it, which is $v_{\sqrt{n},\sqrt{n}}$, anonymously. That is, the destination node $v_d = v_{\sqrt{n},\sqrt{n}}$. To bound this communication-cost from above, we consider the path from the client to the anonymity-set s_j, to which v_d belongs, and back to the client via the cloud. Our routing algorithm splits the route to v_d into three parts, as in section 4.2.2. Intuitively, the total incurred communication-cost is equal to the sum of query and response routing cost along these three parts of the route, plus the data collection piggyback cost. We claim that this total communication-cost is at worst in $O(d^2)$.

To see why, first recall from section 4.2.2 that our routing algorithm relies on the first-node v_1 and the last-node $v_{|s_j|}$ of the anonymity-set s_j to which $v_d = v_{\sqrt{n},\sqrt{n}}$ belongs. The first part routes the query from root-node $v_{1,1}$ to v_1. Since the diameter of the graph d is the longest shortest path between any pair of nodes, the length of the first part of the route is in $O(d)$. Similar argument is valid for the third part which routes the piggybacked responses from $v_{|s_j|}$ to $v_{1,1}$. Thus, the length of the third part is in $O(d)$. Additionally, the length of the second part is also in $O(d)$. To reason about that, recall that the length of this part is equal to the size of the anonymity-set $|s_j|$ to which the node $v_{\sqrt{n},\sqrt{n}}$ belongs. In DAS, we have $|s| \in [k, 2k-1]$. Thus, $|s| = \Theta(k)$, and by assumption we have $k \in [1, \Theta(d)]$, hence the result follows. Now, since the length each of theses three parts is in $O(d)$, the sum of them is also in $O(d)$. By lemmas 2, the first, second, and third parts incur a worst-case routing cost of $O(d^2)$.

Next, we shall examine the piggyback cost. The second part of the route is used to collect responses from all nodes in the anonymity-set s_j in a piggyback manner. By lemma 3, the incurred communication-cost from this part is in $\Theta(k^2)$. But, by assumption, $k \in [1, \Theta(d)]$, thus the worst-case data collection cost is also in $O(d^2)$. This collected data of size$O(k)$ is forwarded along the third part of the route, whose length is in $O(d)$, back to the client. This incurs a worst-case cost of $O(kd)$. Since $k \in [1, \Theta(d)]$, the incurred worst-case cost is in $O(d^2)$. Consequently, The total communication-cost is at worst: $c_w = O(d^2) + O(d^2) + O(d^2) = O(d^2)$.

Necessity: assume otherwise for the purpose of contradiction. That is, there exists a secure k-anonymous query scheme that achieves a level-of-anonymity $k \in [1, \Theta(d)]$, and requires a communication-cost of $o(d^2)$ in the worst case. We observe that the worst-case communication-cost is incurred when the client seeks to query $v_{\sqrt{n},\sqrt{n}}$, the farthest node from it. Thus, its worst-case communication-cost is bounded from below by the communication-cost of sending at least 1 bit to node $v_{\sqrt{n},\sqrt{n}}$ and back to the client via cloud. This cost represents the least communication-cost that is needed by any anonymity scheme to perform its goal since it is the minimum cost needed to communicate with node $v_{\sqrt{n},\sqrt{n}}$ without offering any anonymity, i.e. $k = 1$. It is easy to see that this communication-cost is in $\Omega(d^2)$. First, consider the Manhattan distance from the root-node $v_{1,1}$ to $v_{\sqrt{n},\sqrt{n}}$ and back. The distance on each direction is linear in the diameter of the graph d. Then the result, which is $c_w = \Omega(d^2)$, follows from lemma 2.

□

Theorem 9. *To achieve secure query $k - anonymity$ of $k \in [\Theta(d), n]$ in a source-routed square grid WSN of size n whose diameter is d, a worst-case cost of $c_w = O(k^2)$ is sufficient, and $c_w = \Omega(k^2)$ is necessary.*

Proof. Sufficiency: we prove this part by construction. That is, we provide a secure k-anonymous query scheme, which is DAS, and show that a worst-case cost of $O(k^2)$ is sufficient to achieve a level-of-anonymity of $k \in [\Theta(d), n]$. As in the proof of Theorem 8, we drive the worst-case communication-cost by setting the destination node $v_d = v_{\sqrt{n},\sqrt{n}}$ in our routing algorithm, i.e. Algorithm 2. To articulate an upper bound for the routing cost, we consider the three parts of the route to s_j, which is the anonymity-set to which $v_d = v_{\sqrt{n},\sqrt{n}}$ belongs. The first and third are shortest paths between two points in the graph. Therefore, each of them is in $O(d)$. Now by assumption, we have $k \in [\Theta(d), n]$. That is $k \geq \Theta(d)$. Thus each of the first and second path length is in $O(k)$. It is easy to see that the length of the second part is also in $O(k)$ since it is equal to $|s_j|$, which is in $\Theta(k)$ for DAS. As a consequence, the length of the three parts is in $O(k)$. Thus the the routing cost is at worst in $O(k^2)$ by By lemmas 2. Likewise, the data collection piggyback cost is in $O(k^2)$ by lemma 3. Furthermore, the collected data of size $O(k)$ is forwarded along the third part of the route, which is of length $O(d)$, back to the client. This incurs a worst-case cost of $O(kd)$, but since $k \in [\Theta(d), n]$ by assumption, hence the incurred worst-case cost is in $O(k^2)$. Therefore, the total communication-cost, which is the sum of the routing and data collection cost, is in $O(k^2)$.

Necessity: For the purpose of contradiction, assume that there exists a secure k-anonymous query scheme that achieves a level-of-anonymity $k \in [\Theta(d), n]$, and requires a communication-cost of $o(k^2)$ in the worst case. However, such a scheme must be able to query and collect a response from the worst-case destination node $v_d = v_{\sqrt{n},\sqrt{n}}$ anonymously. As mentioned in Chapter 4, this entails that the scheme must be able to query and collect responses from all nodes in the anonymity-set s_j to which the destination node $v_d = v_{\sqrt{n},\sqrt{n}}$ belongs. Consequently, the source route must include at least $|s_j|$ nodes, and at least 1 bit must be sent to each one of them. By Lemma 2, The cost of this communications is in $\Theta(|s_j|^2)$. Hence, the scheme incurs a necessary communication-cost of $\Omega(|s_j|^2)$. We obtain easily a desired contradiction by applying corollary 2 which yields that the necessary communication-cost is in $\Omega(k^2)$.

□

5.3 Query Path Length

The proceeding analysis deals with the trade-off between query-anonymity and communication overhead cost which is the main focus of our work. Another useful result we establish in Theorem 10 related to the asymptotic bounds on query and response path length. Its importance stems from the fact that it is an indicative of the propagation time delay caused by an anonymity scheme in order to achieve various values of level-of-anonymity k. In Chapter 6, we use the path length as a performance index to compare different configurations of the secure k-anonymous query scheme.

Theorem 10. *To achieve secure query k-anonymity of $k \in [1, \Theta(d)]$ in a source-routed square grid WSN of size n whose diameter is d, the query and response path length measured in number of hops is in $O(d)$ in the average-case and worst-case, and to achieve $k \in [\Theta(d), n]$, the path length is in $O(k)$ in the average-case and worst-case.*

Proof. The proof is straightforward once we recall that the hops along each query and response path of a secure k-anonymous query scheme are of two types namely, the routing-hops, and the collection-hops. The routing-hops are used to forward the query to the firs-node of the anonymity-set, and to collect the responses from the last-node of it. Therefore, the routing-hops are in $O(d)$. On the other hand, the collection hops are the anonymity-set nodes whose responses are collected. Thus, the collection-hops are in $O(k)$ because the anonymity-set size is in $O(k)$ by corollary 2. Now, when d is relatively large compared to k, i.e. $k \in [1, \Theta(d)]$, the routing-hops dominates the collection-hops so that the whole path length is in $O(d)$, and vice versa. That is for $k \in [\Theta(d), n]$, the collection-hops dominates resulting in the path length is in $O(k)$.

□

Chapter 6

Performance-Anonymity Trade-Offs Evaluation

The proposed secure k-anonymous query scheme, DAS, was examined via extensive simulation. We ensure all the targeted variables, such as average-case, and worst-case cost, path length, and location-anonymity, are systematically measured with respect to the level-of-anonymity k, WSN network size n, and diameter d. We implemented all of our algorithms and querying protocol in a large-scale network environment on a dedicated 2.6 GHz Intel Core $i7$ x86_64 bit machine running OS X 10.11.6 El Capitan. In essence, we measured the following:

- How the anonymity scheme performance changes with the level-of-anonymity k for various values of WSN network diameter d, and
- How the anonymity scheme performance changes with WSN network diameter d for various values of level-of-anonymity k.

The query packet is structured as follows: 1 byte for the client ID, 10 bytes for MAC header, 1 byte for the action taken by each node to be either froward or append response then forward, 2 bytes for the node IDs. Each query data and response data is modeled to be constant at 32 bytes. To evaluate energy consumption, we adopt the MEMSIC TelosB mote platform [8] where each node is composed of a TI MPS 430 microcontroller and CC2420 RF transceiver that consumes 1.8 microjoule per byte for reception and 2.1 microjoule per byte for transmission.

In section 6.1, we introduce our cost-benefit metric ROI, and location-anonymity metrics r and d_{max} that prove to be beneficial in our various performance-anonymity trade-offs analysis presented in section 6.2.

6.1 Performance Metrics

In addition to various performance anonymity trade-offs, we are interested in measuring the offered level-of-anonymity k (the benefit) relative to the communication-cost c (cost) of implementing the proposed scheme, as well as evaluating the achieved location-anonymity. For the cost-benefit analysis, the metric of Return-On-Investment (ROI) is introduced, where $ROI = \frac{k}{c}$. In essence, ROI measures the return on investment relative to the cost of investment. ROI is taken for both the average and worst-case communication-cost evaluation.

For the location-anonymity, two formal metrics are adopted to measure the average and maximum achieved anonymity. The first metric is the *radius of cloaking area* of a specific level-of-anonymity k that is modelled as an equivalent circular area. Note that the cloaking area is the amount of space within which the true destination of a query is not identifiable due to the existence of its anonymity-set members, whereas the cloaking distance of a node, denoted as d_c, is the distance between it and any another node in its anonymity-set [93]. The radius r measures the average cloaking distance of the centroid node over all anonymity-sets. Note that the centroid node of the anonymity-set s_j is the node v_i with the minimum sum of cloaking distances $Sum_i d_c$ to all other nodes in s_j, i.e., $\min_{v_i \in s_j}(Sum_i d_c)$. Thus, we compute the radius r as the minimum value of the average of the cloaking distances of the centroid over all anonymity-sets for a specific k, i.e.,

$$r = \min_{s_j \in \mathbb{S}} \left(\frac{\min_{v_i \in s_j}(Sum_i d_c)}{k-1} \right), i \in [1, |s_j|], j \in [1, \lfloor n/k \rfloor]$$

Now, we show how to obtain the second metric, i.e. the maximum cloaking distance d_{max}. We first find the maximum cloaking distance d_{max_i} for each node v_i in an anonymity-set s_j for a certain level-of-anonymity k. Since each node in the anonymity-set is a possible true destination, then we apply the concept of weakest link to select the minimum of the found maximum cloaking distances for each anonymity-set [83]. Eventually, we choose the weakest link again by taking the minimum over all of the anonymity-sets which results in the final maximum cloaking distance d_{max}. That is,

$$d_{max} = \min_{s_j \in \mathbb{S}} (\min_{v_i \in s_j} (d_{max_i})), i \in [1, |s_j|], j \in [1, \lfloor n/k \rfloor]$$

We specify in Algorithm 4 how to compute r and d_{max}. In addition, a simplified example in which $k = n = 9$, i.e. for a single anonymity-set case, is shown pictorially in Fig. 6.1.

We are interested in studying the limiting behavior of location-anonymity that is offered by a secure k-anonymous query scheme such as DAS. In Theorem 11 below, we derive asymptotic upper-bounds on d_{max}, and r.

Theorem 11. *Given a connected undirected graph $G = (V, E)$ of n vertices whose diameter*

Algorithm 4 The Location-Anonymity Metrics Algorithm

```
Input: A partition of $\sqrt{n} \times \sqrt{n}$ square grid connected undirected graph $G = (V, E)$ into
    disjoint anonymity-sets $\mathbb{S}$ for a desired level-of-anonymity $k$.
Output: $r$, and $d_{max}$ location-anonymity metric.
$n \leftarrow |V|$
Initialize each of $D, M, M', A$ and $A'$ to an empty list
foreach anonymity-set $s \in \mathbb{S}$ do
    foreach node $v \in s$ do
        $s' \leftarrow s$ after node $v$ is removed
        foreach node $v' \in s'$ do
            Append the distance from $v$ to $v'$ to $D$   /* D stores the cloaking distances of
            node $v$. */
        end for
        Append $Max(D)$ to $M$
        Append $Sum(D)$ to $A$
    end for
    Append $Min(M)$ to $M'$
    Append $Min(A)/(k-1)$ to $A'$
end for
$d_{max} \leftarrow Min(M')$
$r \leftarrow Min(A')$
return $r$, $d_{max}$
```

is d, both of the location-anonymity metrics d_{max} and r are in $O(d)$ for DAS secure k-anonymous query scheme that achieves a level-of-anonymity $k \in [\Theta(d), n]$, and in $O(k)$ for secure k-anonymous query scheme that achieves a level-of-anonymity $k \in [1, \Theta(d)]$.

Proof. The key to the proof is to note that the each of the cloaking distances of a node in an anonymity-set may not exceed $O(k)$ and $O(d)$. To see why, we reason as follows. The anonymity-set size is $|s| \in [k, 2k-2]$, thus we have $|s| \in O(k)$ which is consistent with Corollary 2. Recall that we measure the distance, including the cloaking distance d_c, between two nodes as the length of the shortest path connecting them. Since the nodes in the DAS anonymity-set are adjacent then we have $d_c \in O(k)$. d_c is also in $O(d)$ because d is the longest shortest path between any pair of nodes by definition. To have a tight bound on d_c, for relatively large k compared to d, i.e. $k \in [\Theta(d), n]$, we bound d_c from above by $O(d)$, and vice versa. That is, for relatively large d compared to k, i.e. $k \in [1, \Theta(d)]$, we bound d_c from above by $O(k)$. Now, since d_{max} is one of the cloaking distances, and r acts as an average of the cloaking distances then the desired result holds. □

Now, since the square grid is a connected undirected graph, the assertion established in Theorem 11 is applicable to it. In section 6.2, we conduct an exact experimental analysis of the location-anonymity metrics to investigate its limiting behavior for various k and d.

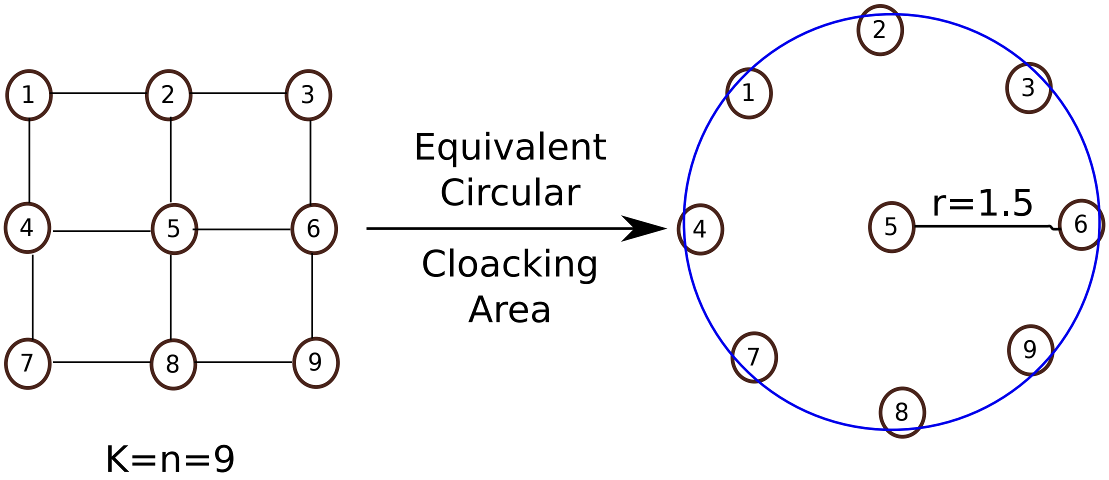


Figure 6.1: A simplified example of creating the equivalent circular cloaking area, and computing its radius r for the case of single anonymity-set, i.e. $k = n = 9$. We have the sum of cloaking distances for each node as follows: $Sum_1 d_c = 1{+}2{+}1{+}2{+}3{+}2{+}3{+}4 = 18$, $Sum_2 d_c = 15$, $Sum_3 d_c = 18$, $Sum_4 d_c = 15$, $Sum_5 d_c = 12$, $Sum_6 d_c$=15, $Sum_7 d_c$=18, $Sum_8 d_c = 15$, $Sum_9 d_c = 18$. We conclude that node 5 is the centroid of the cloaking area since it has the minimum sum of cloaking distance. Now, we compute the average cloaking distance of node 5, the centroid to be $12/(9-1) = 1.5$. It is only one value in this case since we have a single anonymity-set. Thus, there is no need to do the last step namely, selecting the minimum average of cloaking distances of centroids over all anonymity-sets. In terms of d_{max} calculation for the same example, the maximum clocking distances of each node in the anonymity-set are as follows:$d_{max_1=4}$, $d_{max_2=3}$, $d_{max_3=4}$, $d_{max_4=3}$, $d_{max_5=2}$, $d_{max_6=3}$, $d_{max_7=4}$, $d_{max_8=3}$, $d_{max_9=4}$. Thus, the lowest value is 2 which belongs to node 5. This is also the final value of d_{max} since we have a single anonymity-set.

6.2 Evaluation Results

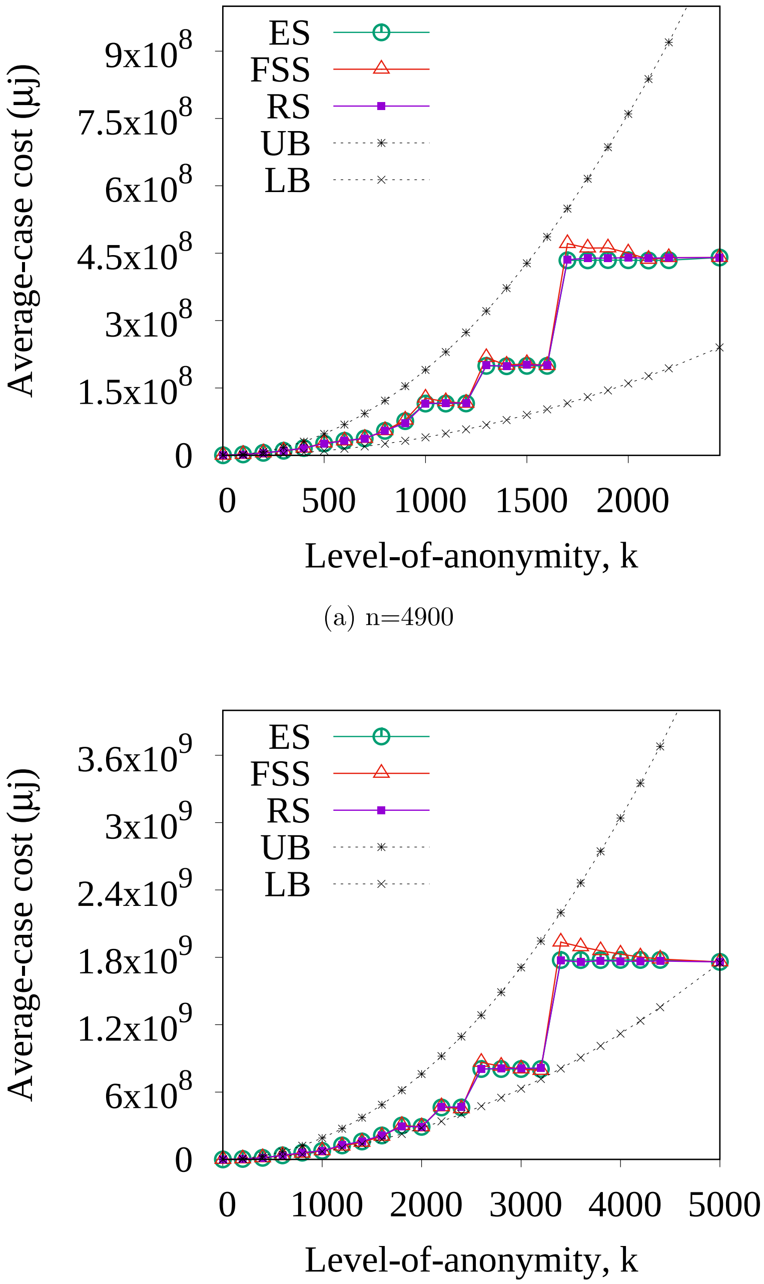
The main goal of our experiments was to validate our asymptotic analysis presented in section 5.2. The crucial observation we can make is that our exact average and worst-cost analysis of the achieved query-anonymity reasserts our asymptotic bounds for large scale of WSN network size and simulation settings.

Our asymptotic cost analysis, presented in section 5.2, indicated that the behavior of the secure k-anonymous query scheme is shaped by two main variables, namely the level-of-anonymity k and the network diameter d that is further a function of WSN size n. Thus two sets of experiments are conducted as in the following sections in order to measure the performance of the anonymity scheme after changing one of these variables separately. Due to the large size of data that we collected, we will be discussing below typical representative results and performance trends.

6.2.1 Level-of-Anonymity k Impact

Average and Worst-Case Communication-Cost

We first measured the average communication-cost c_a incurred by DAS to achieve various level-of anonymity $k \in [1, n]$ based on a specific anonymity-set construction method FSS, ES, or RS, respectively. Since the network diameter d of a square grid graph is equal to $2(\sqrt{n} - 1)$, changing n implies change of d. For more reliable results, our experiment is repeated in several WSN networks of sizes up to $n = 10000$, that is of diameter up to 198. In addition, we averaged the RS results over 100 random runs. We depict the obtained results for larger network sizes in Fig. 6.2.



(a) n=4900

(b) n=10000

Figure 6.2: The average communication-cost with varying k.

It is observed that there exists a value of $k \in \Theta(d)$, which we call $k_0 = c_0 d$, after which the average communication-cost c_a is bounded from above and below by $O(k^2)$, and $\Omega(k^2)$ respectively, as stated in Theorem 7. The upper bound UB and the lower bound LB serve as sufficiency and necessary conditions in Theorem 7, respectively. These bounds provide a sound empirical evidence in support of our asymptotic analysis.

In the second column of Table 6.1, we summarize the estimated values of the constants for each of the above asymptotic relationships.

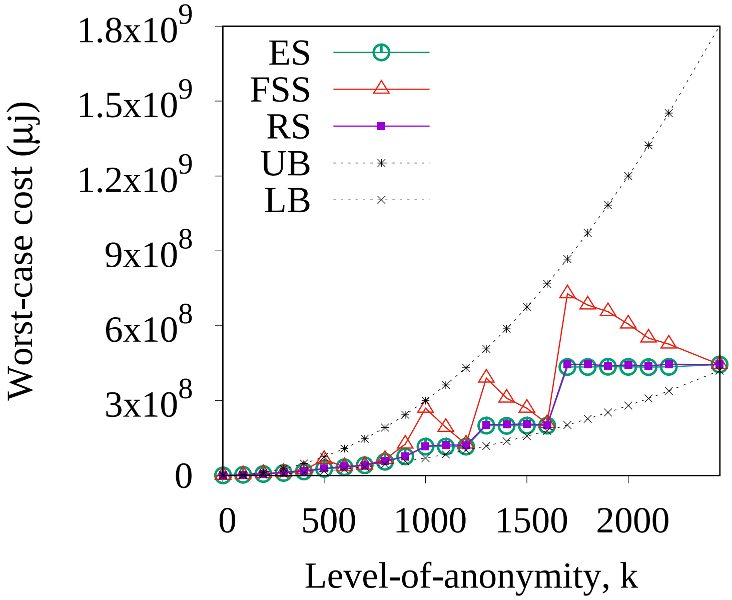
Compared to DC-Net which occurs when $k = n$ (i.e. $k = 4900$ or $k = 10000$), DAS provided a more flexible range of levels-of-anonymity with lower average-case communication-cost. Indeed the incurred communication by DC-Net scheme was so high (1.6×10^9 microjoule for $n = 4900$, and 7×10^9 microjoule for $n = 10000$) that we did not include in Fig. 6.2 to be able to compare the little differences among different anonymity-sets construction methods. In terms of comparing the different anonymity-sets construction methods, we observed that they are close to each other with respect to their incurred average communication-cost to achieve various values of k. It is expected that the FSS takes less energy because it tends to have the whole remainder in the first set close to the root node eliminating the first and third part of the source-route, as described in Algorithm 2. However, it was quiet surprising that FSS was outperformed consistently by the ES and RS at some points. These points occur when the remainder of n/k is at its local maximum value. A reason for that must be related to the way that the piggy-back strategy collects sensor readings which incurs a cost that grows with the remainder of n/k. Additionally, the differences in energy consumption become less evident as the remainder approaches zero. This is what gives the curves their stair-like shapes.

Fig. 6.3 shows the results of worst-case communication cost. Similar to that of the average-case above, the worst-case cost is bounded by k^2 after some value of $k \in \Theta(d)$, namely $k_0 = c_0 d$ as establish in Theorem 9. For better comparison, we used the same set of k values for all of our tests in this section. Intuitively, the estimated value of the upper bound asymptotic constant c_2 is significantly higher than that of the average-case since we are changing k along the same range of values. Interestingly enough, $c_0 = 1$ is the same value as that of the average-case. This highlights the fact that $k_0 = c_0 d = d$ is the inflection point above which, i.e. $k \in [d, n]$ or $k \gg d$, the quadratic term that dominates the asymptotic behavior of the average and worst-case cost is k^2. We list these findings in Table 6.1.

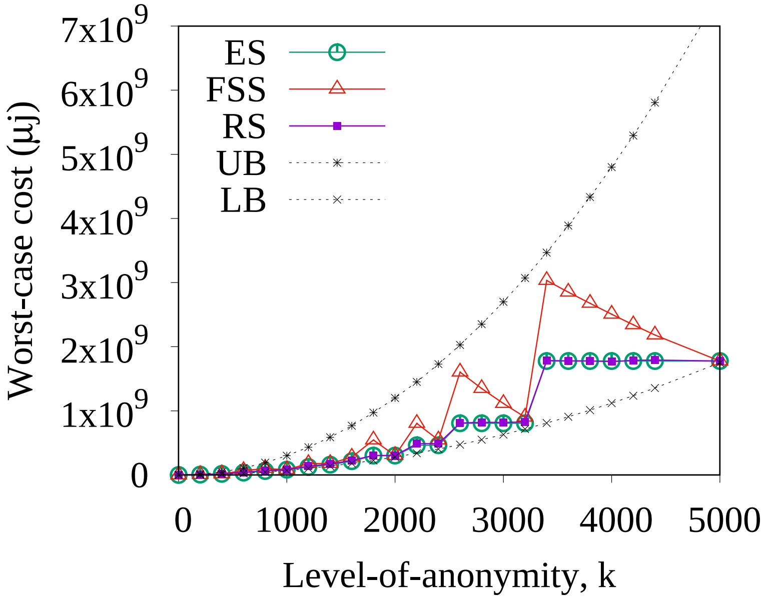
Table 6.1: Values of Asymptotic Constants when $k \gg d$

Asymptotic constant	Average-case analysis	Worst-case analysis
c_0 of $\Theta(d)$	1	1
c_1 of $\Omega(k^2)$	40	70
c_2 of $O(k^2)$	190	300

Compared to DC-Net which occurs when $k = n$ (i.e. $k = 4900$ or $k = 10000$), DAS provided a more flexible range of levels-of-anonymity with lower worst-case communication-cost. Again, the incurred communication by DC-Net scheme was so high (1.6×10^9 microjoule for $n = 4900$, and 7×10^9 microjoule for $n = 10000$) that we did not include in Fig. 6.2 to be able to compare the little differences among different anonymity-sets construction methods. Similar observations are gained as that of the average-case above where the worst-case cost of FSS was the highest, the Random-Spread performs in between the FSS and the ES, but closer to the later due to its random assignment of the remaining nodes to different sets.



(a) n=4900



(b) n=10000

Figure 6.3: The worst-case communication-cost with varying k .

Cost-Benefit Analysis

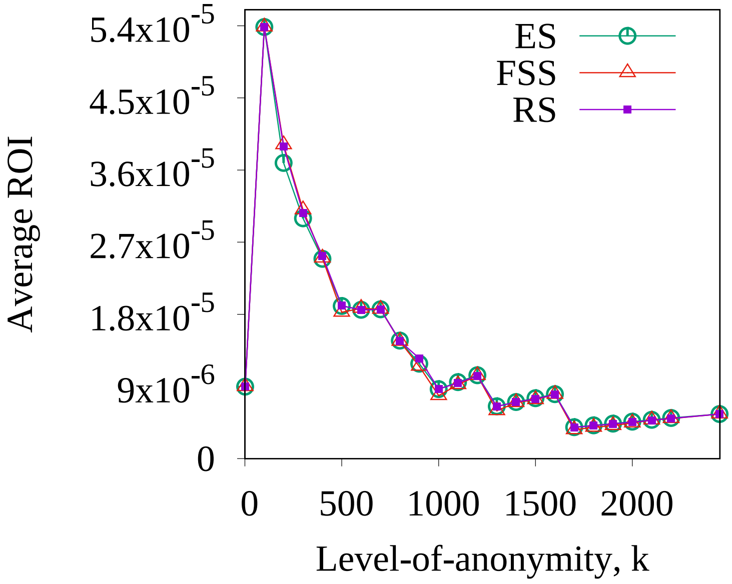
There is something specific and enriching that we discover by looking into the curves and data traces of ROI for variable level-of-anonymity k as shown in Fig. 6.4, and 6.5. The ROI curves were shaped like a bell around the inflection point $k_0 = d$ for both the average and worst-case cost, and all the anonymity-sets construction methods. That is, the ROI increased with k till the inflection point $k_0 = d$ after which it decreased significantly with k. From practical viewpoint, this indicates that the ROI relative to the cost is at its highest values around the inflection point $k_0 = d$.

We argue as follows. First recall from the asymptotic analysis, which is presented in section 5.2, that the communication-cost c consists of two main components, one grows in proportion to k^2 and the other is in d^2. The first component increases with the anonymity-set size $|s|$ since $|s| \in [k, 2k-1]$. On the other hand, the second is due to the cost of forwarding a query to the first-node of anonymity-set, and returning collected responses from the last-node of the anonymity-set. Since we fixed the network size n during this part of the experiment, the length of the forward and return paths shrinks with every increment in k. It follows that this component decreases with k. Now, for the range of $k < k_0$, The increment in the first component is marginal and balanced by the decrease in the second leaving the communication-cost almost constant. Thus, $ROI = k/c$ increases with k for $k < k0$. Nevertheless, when k exceeds the inflection point $k_0 = d$, the first component k^2 dominates the second which results in a substantial decrease in ROI as the value of k grows.

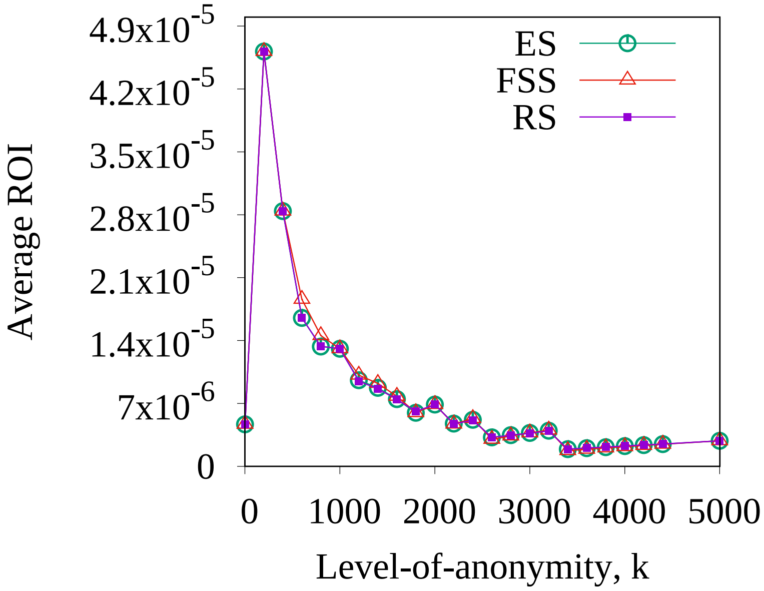
Compared to DC-Net which occurs when $k = n$ (i.e. $k = 4900$ or $k = 10000$), DAS provided better ROI in most of the cases because of the high incurred communication-cost of DC-Net whose ROI was fixed at 1.4×10^{-6} and 2.9×10^{-6} for $n = 10000$ and $n = 4900$

respectively in the average and worst-case.

By comparing the ROI data traces of different anonymity-sets construction method, we made comparable observations to that of the average and worst-case analysis mentioned in the proceeding paragraphs. That is, the values of the ROI under the three methods were close to each other. However, the ROI of FSS was the lowest in most of the time as k increased due to the growing size of the first anonymity-set. This becomes more evident when the worst-case cost is used to compute the ROI as shown in Fig. 6.5. Additionally, the ROI of RS was between the FSS and the ES.

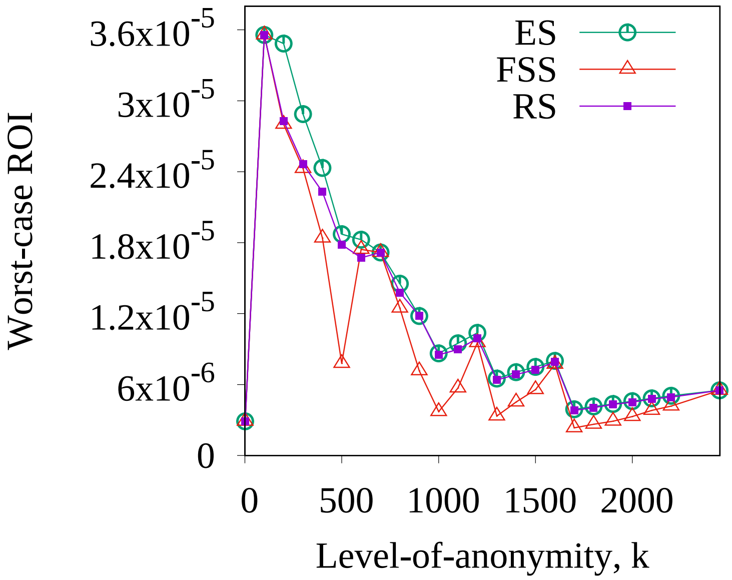


(a) n=4900

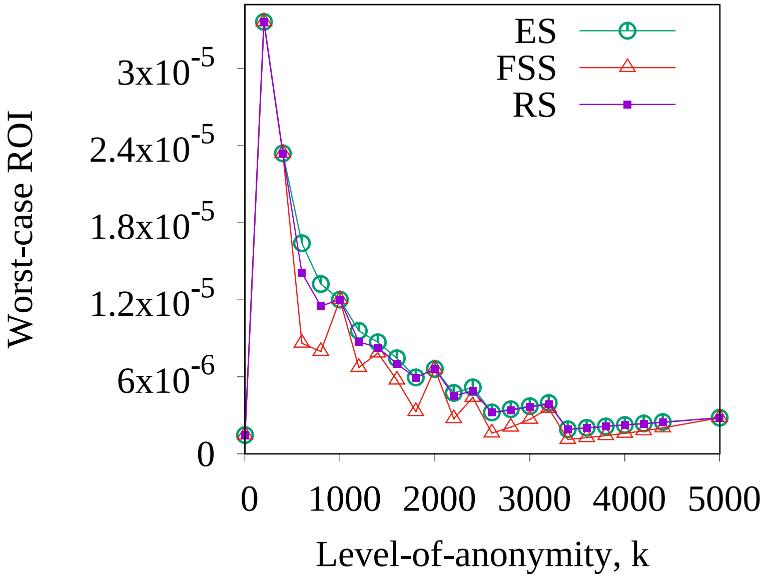


(b) n=10000

Figure 6.4: The cost-benefit metric using the average communication-cost ROI_{av} with varying k .



(a) n=4900



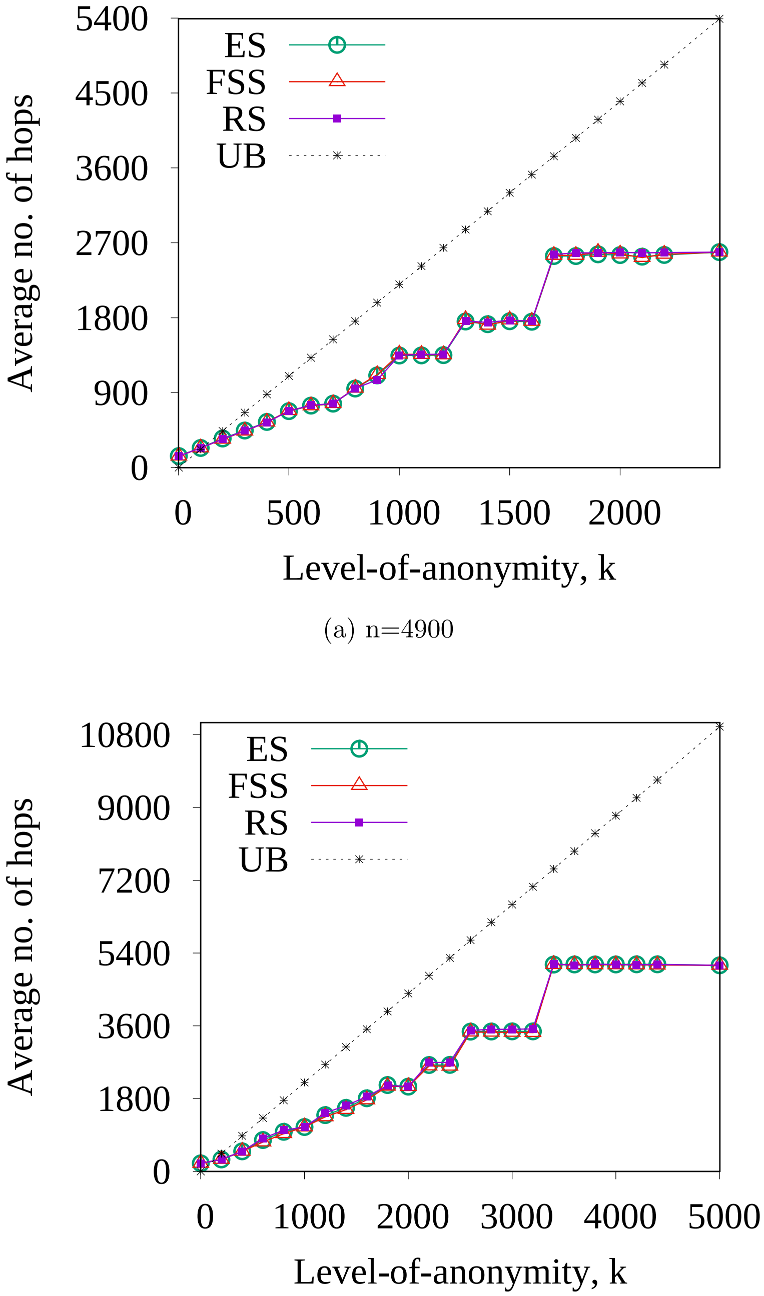
(b) n=10000

Figure 6.5: The cost-benefit metric using the worst-case communication-cost ROI_w with varying k

Path Length

Furthermore, we found that the total number of hops along the query and response path are in $O(k)$ for $k > d$ for both the average and worst-case, as shown in Fig. 6.6 and 6.7, regarding various n and anonymity-sets construction methods. This confirms our theoretical assertions in Theorem 10.

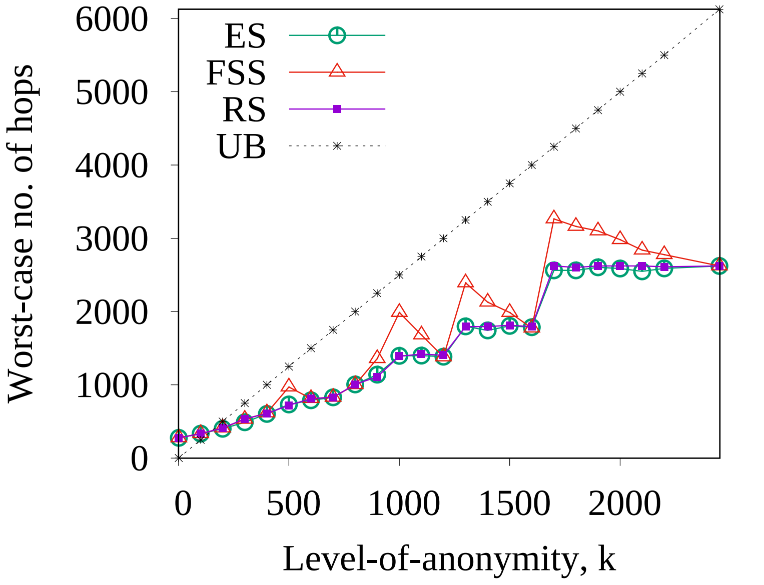
As expected, DC-Net which occurs when $k = n$ (i.e. $k = 4900$ or $k = 10000$) produced longer path length (4969 for $n = 4900$, and 10099 for $n = 10000$) in the average and worst-case than DAS. The path length is further used as an index to compare the performance of the different anonymity-sets construction methods. We observed that all of them performed similarly in the average-case, which follows our intuition since the different anonymity-sets construction methods spread the same total number of hops in various ways. Thus, they produced comparable total number of traversed hops to query all the anonymity-sets and collect their responses. Nevertheless, the worst-case path length generated by the FSS is found longer than that by ES and RS whenever there is a remainder out of n/k. To see why, recall that the hops along the query and response path are of two types: routing-hops which increase with d, and collection-hops which increase with k. Now since the ES attempts to distribute the remaining nodes out of n/k equally on all of the anonymity-sets, the worst-case collection hops grows steadily with k. On the contrary, the FSS adds all the remaining nodes to the first set. Thus, the worst-case number of hops occurs at the first set because it has the maximum number of collection hops. Since d was fixed during this set of experiments, the collection hops dominated the routing hops when k grows larger than d. This resulted in that the worst-case path length of FSS surpassed that of the ES whenever there is a remainder. On the other hand, RS acted closely to the ES because it disseminates the remaining nodes out of n/k randomly on the anonymity-sets which makes it divergent from the FSS case.



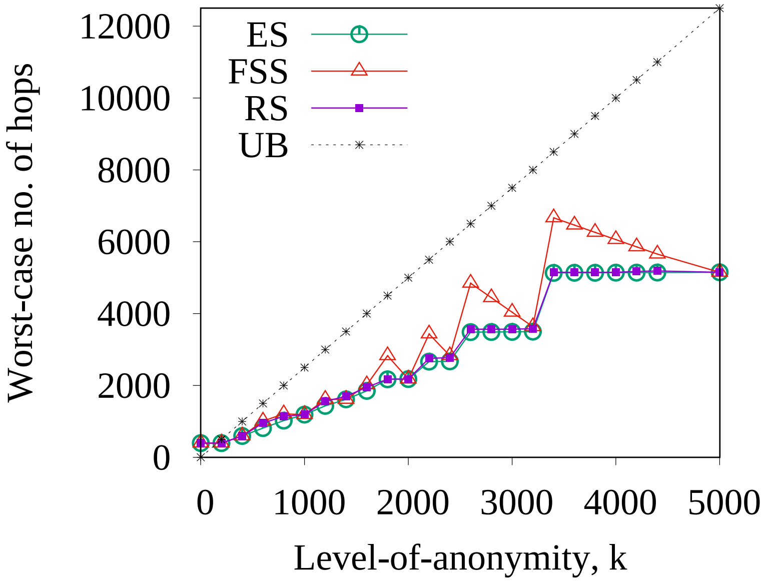
(a) n=4900

(b) n=10000

Figure 6.6: The average number of hops with varying k .



(a) n=4900



(b) n=10000

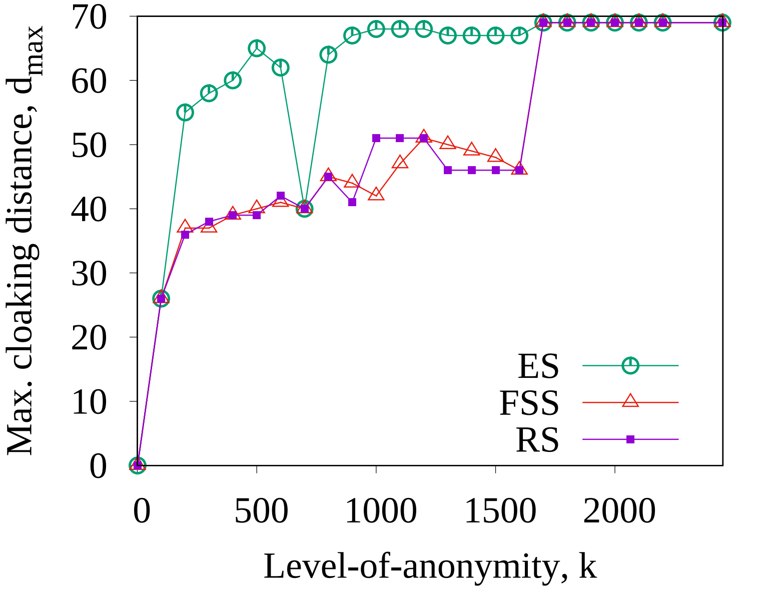
Figure 6.7: The worst-case number of hops with varying k .

Location Anonymity

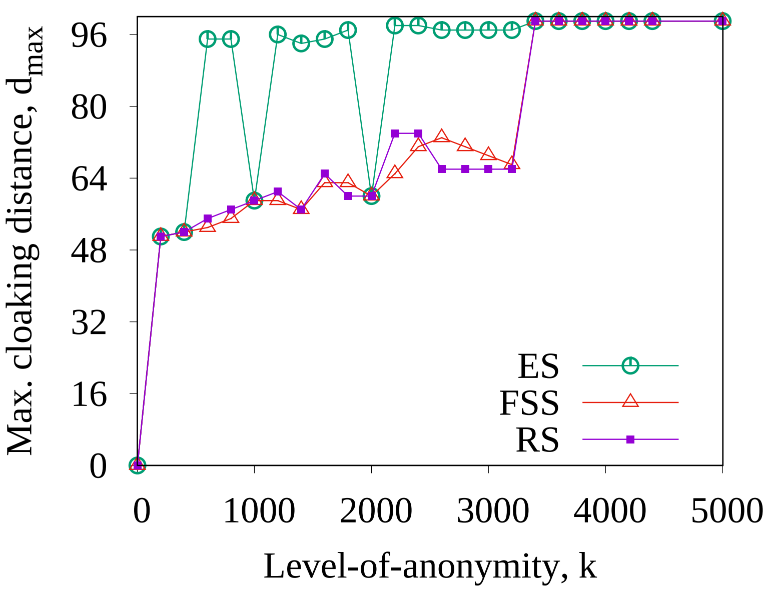
We now evaluate the achieved location-anonymity by DAS anonymity scheme using the maximum cloaking distance d_{max}, and the radius of cloaking area r metrics that are discussed in section 6.1. As expected, we observed that both the d_{max} and r grew with k because of the increase in the size of the anonymity-set. When k is relatively large compared to d, the upper bound of d_{max} and r is in $O(d)$ which confirms our analysis in Theorem 11. Specifically, both d_{max} and r values are less than d. We depict our results for various network size n in Fig. 6.8, and 6.9.

The r and d_{max} values of DC-Net $(k = n)$ were fixed at $d_{max} = \sqrt{n}, r = \sqrt{n}/2$ which represents the maximum offered location-anonymity by any scheme, however the incurred communication-cost was also the highest. By analyzing the location-anonymity metrics data which is shown in Fig.s 6.8, and 6.9, we also noted that the ES achieved better location-anonymity compared to FSS whenever there is a remainder out of n/k. Nonetheless, RS stayed between them in most cases. To reason about it, recall that ES attempts to distribute the remainder equally on all the anonymity-sets while FSS adds all the remaining nodes out of n/k to the first anonymity-set. Thus, the anonymity-sets created by FSS are all of the smallest size which is k except for the first set. On the other hand, when the remainder of n/k is greater or equal to n/k, ES yields anonymity-sets all of a size greater than k. Now, both of d_{max} and r adopts the security weakest link concept as discussed in section 6.1 which always selects the anonymity-set with the minimum size to calculate these metrics. Hence, ES is in advantage by having the minimum size of the anonymity-sets greater than that of the FSS when there is a large enough remainder of n/k. In this way, the metrics of ES become higher than that of FSS as an indicative of better offered location-anonymity. Lastly, because of the randomness of RS, it acts between the two extremes.

As stated in section 6.1, whereas d_{max} measures the location-anonymity that is offered by the anonymity scheme, r acts as a benefit-cost measure. In r, the benefit is sum of cloaking distances, and the cost is the level-of-anonymity which is proportional to the size of anonymity-set that provides the location-anonymity. In essence, r is the radius of the equivalent cloaking area, and it gives a figure about how much location-anonymity is provided if a new node is added to the anonymity-set. By looking into the data traces of r, we discovered ES can provide a larger cloaking area than that by FSS but with less cost.

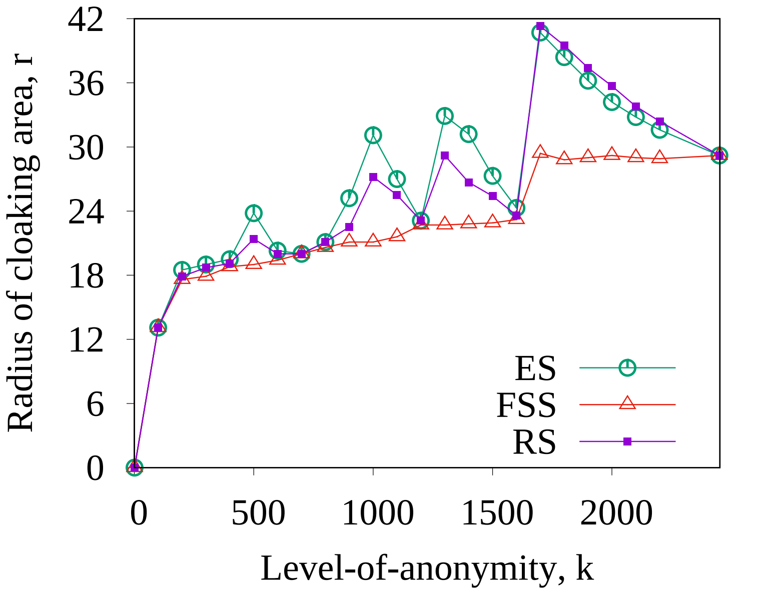


(a) n=4900

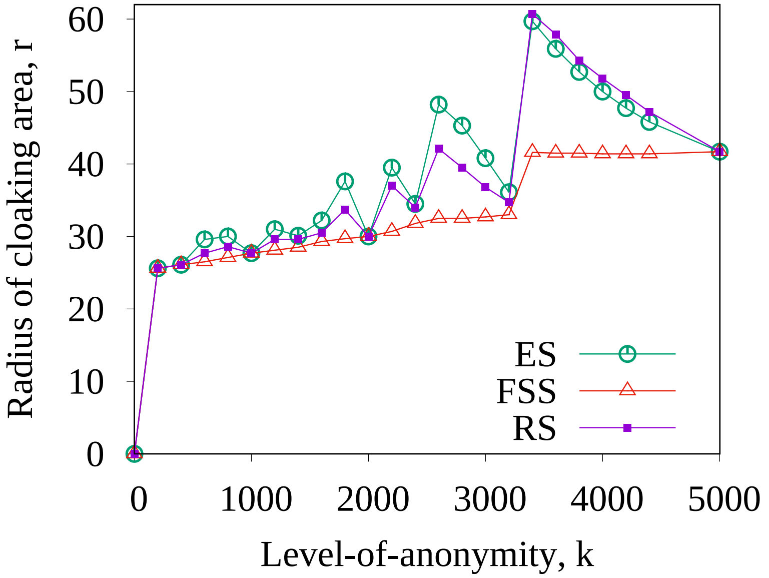


(b) n=10000

Figure 6.8: The maximum cloaking distance metric d_{max} with varying level-of-anonymity k .



(a) n=4900



(b) n=10000

Figure 6.9: The radius of equivalent cloaking area metric r with varying level-of-anonymity k .

6.2.2 Network Diameter d Impact

Average and Worst-Case Communication-Cost

The results of average and worst-case communication-cost obtained by varying d for each value of k are analyzed in this section. As shown in Fig. 6.10 and 6.11, when d exceeds some point $d_0 = k$, the communication-cost in the average and worst-case becomes bounded from above and below by $O(d^2)$ and $\Omega(d^2)$ respectively. This is fully coherent with the result of Theorems 6 and 8. Note that the symmetry of the Big Θ notation entails that $d_0 \in \Theta(k)$ is convertible to $k_0 \in \Theta(d)$ such that the asymptotic constant c_0 of the later equals to the reciprocal of the former.

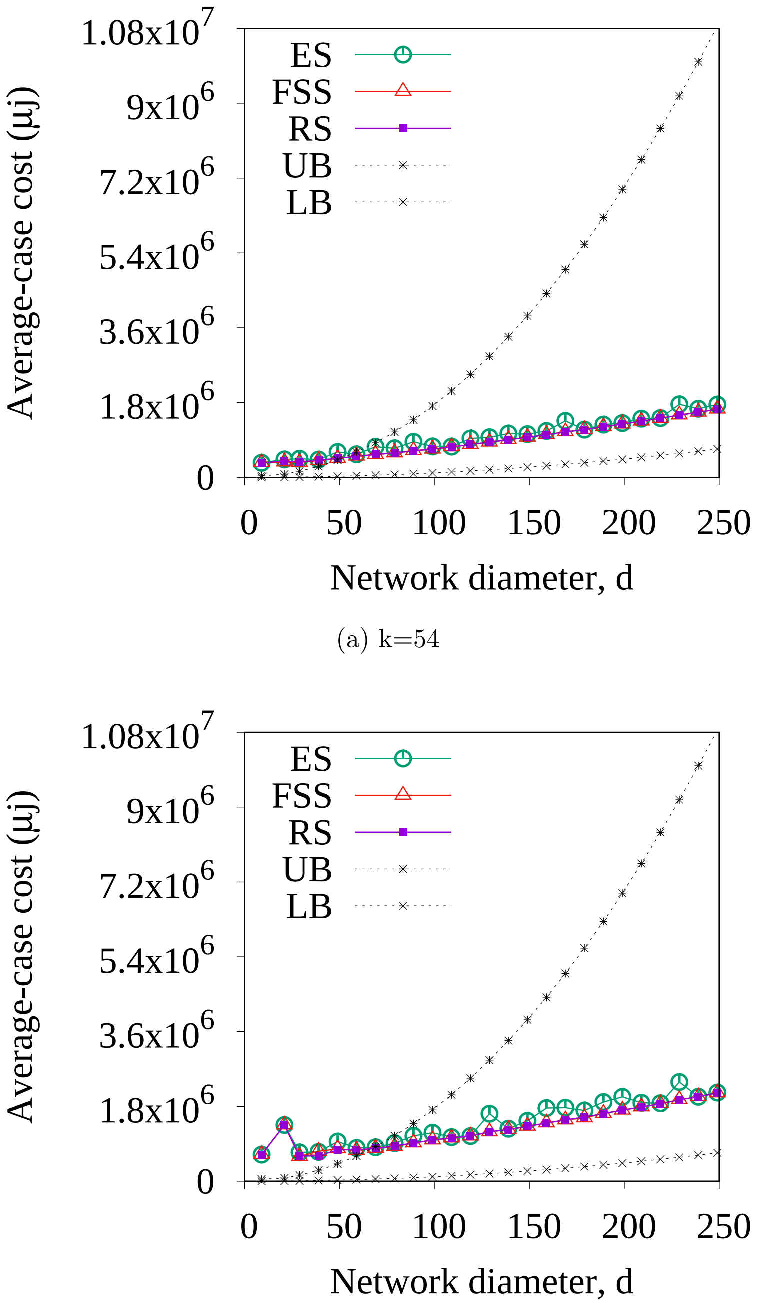
What is significant is that the constant of the asymptotic relation $d_0 \in \Theta(k)$ is seen to be 1. This is exactly the reciprocal of c_0 as recorded for the variable k part of the results, see section 6.2.1. Thus, c_0, which determines the inflection point of the asymptotic relationships, is the same for both parts of the results, namely: variable k, and variable d. This is again fully coherent with the results established in Theorems 6, 7, 8, and 9. Consequently, $k_0 = c_0 d = d$ is a system-wide inflection point at which the asymptotic growth of the communication-cost changes from d^2, when $k \ll d$, into k^2 dominance, when $k \gg d$. The estimated values of other asymptotic constants when d exceeds the inflection point are listed in Table 6.2.

Table 6.2: Values of Asymptotic Constants when $k \ll d$

Asymptotic constant	Average-case analysis	Worst-case analysis
c_0 of $\Theta(d)$	1	1
c_1 of $\Omega(d^2)$	11	30
c_2 of $O(d^2)$	175	250

For the average case, it is observed that ES consumed more energy as d becomes larger compared to k, as shown in Fig. 6.10. Since FSS puts all the remainder of n/k in the first anonymity-set, it is an energy saver with respect to d^2 component of the cost, and a consumer in terms of k^2 component of the cost. The opposite holds for ES. When d surpasses k, the term d^2 becomes more dominant than k^2 component, and vice versa. Consequently, It is more practical to adopt FSS when d exceeds k, as in Fig. 6.10, and Fig. 6.11. However, when k is relatively large compared to d, ES should be preferred, as in Fig. 6.2, and Fig. 6.3.

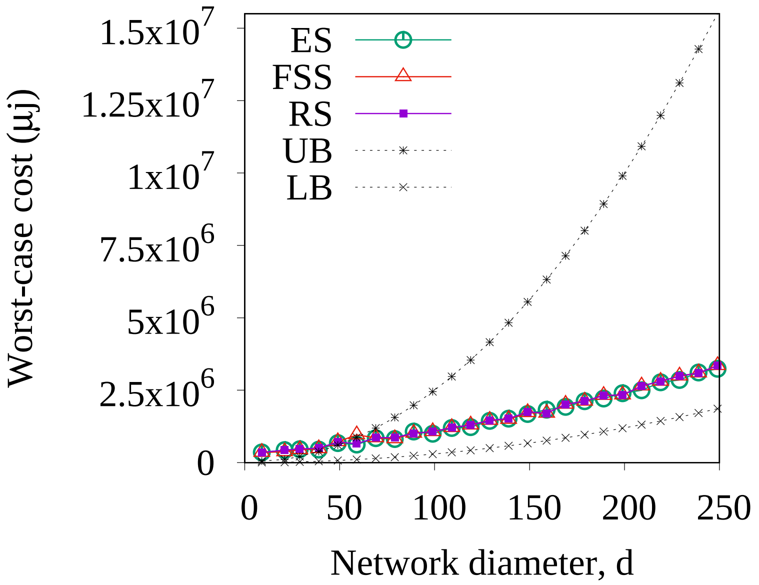
For the worst-case results, as shown in Fig. 6.11, FSS was outperformed by ES specially at large values of k due to the fact that the worst-case cost of FSS occurs in the first anonymity-set which vastly grows with k when there is remainder.



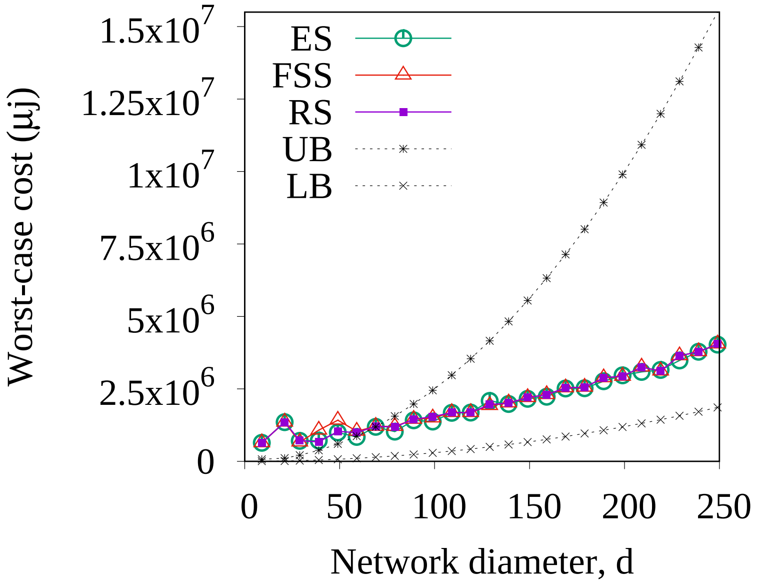
(a) k=54

(b) k=72

Figure 6.10: The average communication-cost with varying d .



(a) k=54



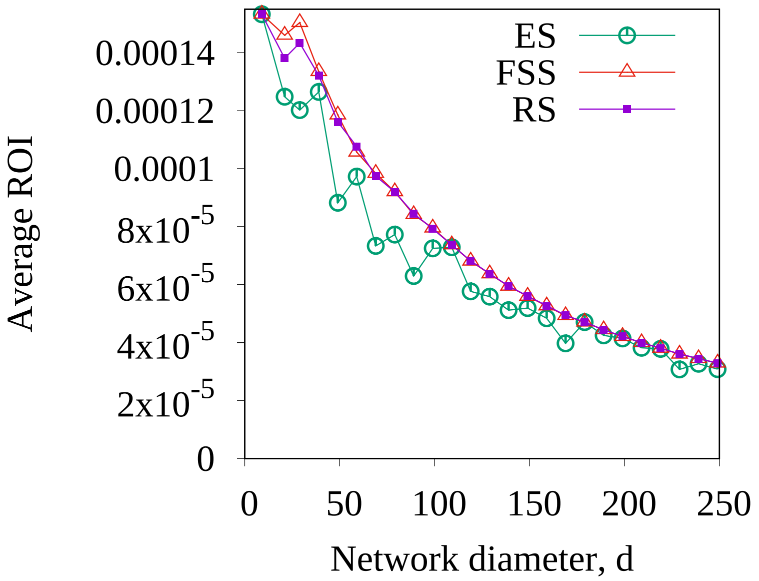
(b) k=72

Figure 6.11: The worst-case communication-cost with varying d .

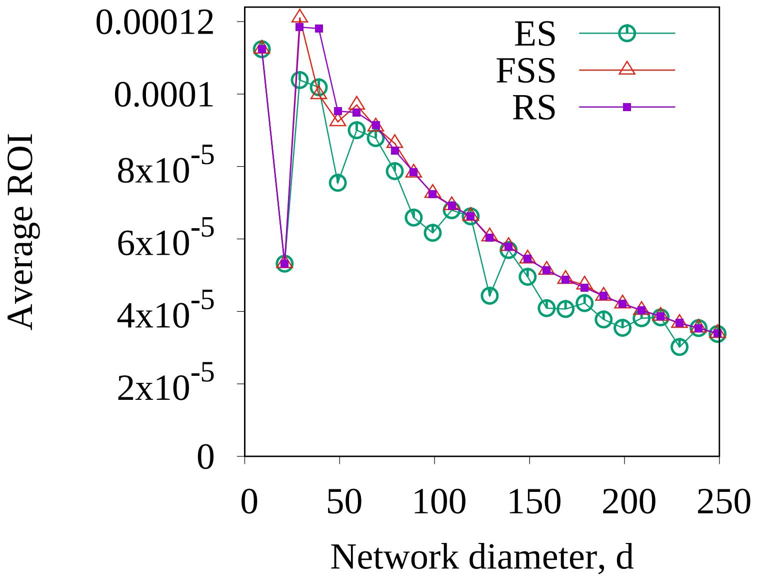
Benefit-Cost Analysis

We have also found that ROI decreased with d for both the average and worst-case cost, as shown in Fig. 6.12, and 6.13, mostly due to the fact of $ROI = k/c$, where c encompasses both the k^2 and d^2 components. Since k is constant in this set of experiments, the sizes of anonymity-sets and the k^2 component of the cost varied slightly due to the mild change in the remainder of n/k. Therefore, approximately the ROI changed inversely with d^2 when d oversteps k. This became more obvious when d grows beyond the inflection point $d_0 = k$ as d^2 dominates k^2.

In analogy with the average and worst-case cost results, the comparison among ROI of different anonymity-sets construction methods of the anonymity scheme reveals similar implications. For the average-case results shown in Fig. 6.12, ES yields the lowest ROI because of its sensitivity to the increment in the d^2 component of the cost. For the worst-case ROI shown in Fig. 6.13, FSS achieves the lowest since the whole remainder was left in the first anonymity-set resulting in the highest worst-case cost, and hence the lowest ROI.

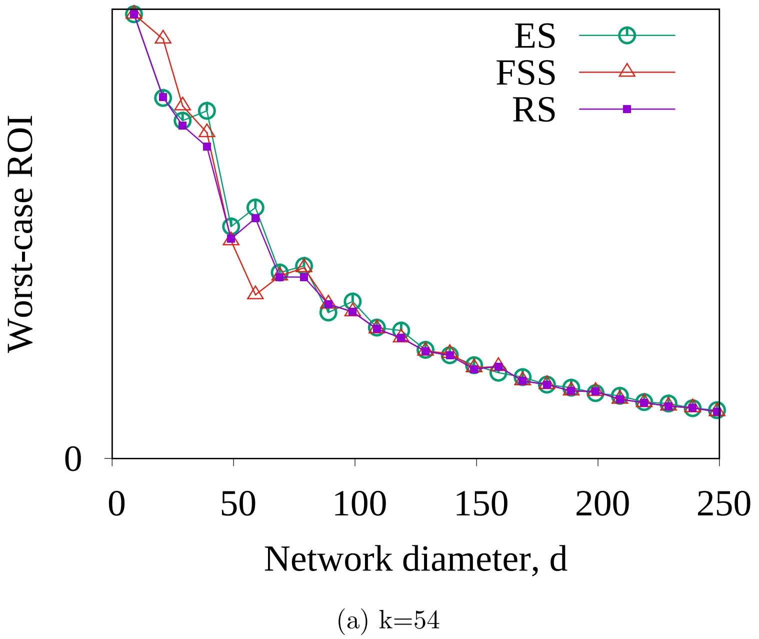


(a) k=54

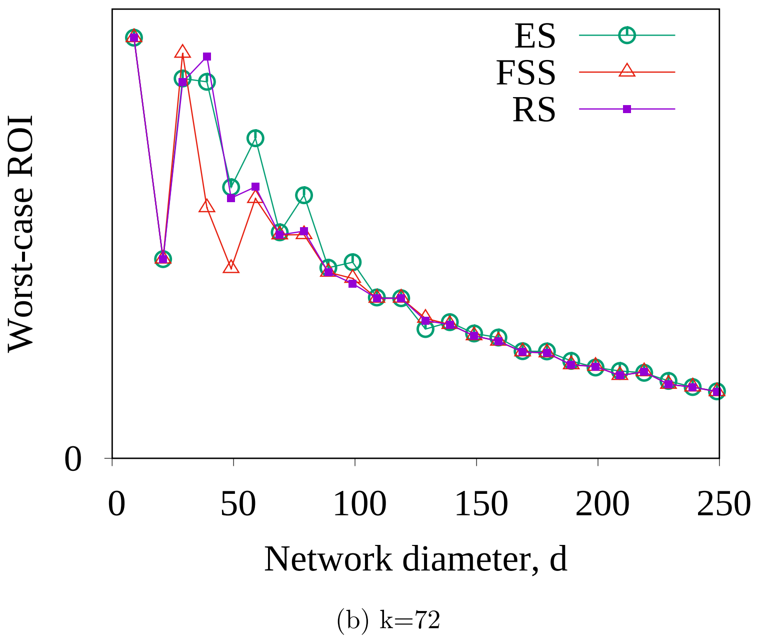


(b) k=72

Figure 6.12: The benefit-cost metric using the average communication-cost ROI_{av} with varying d .



(a) k=54



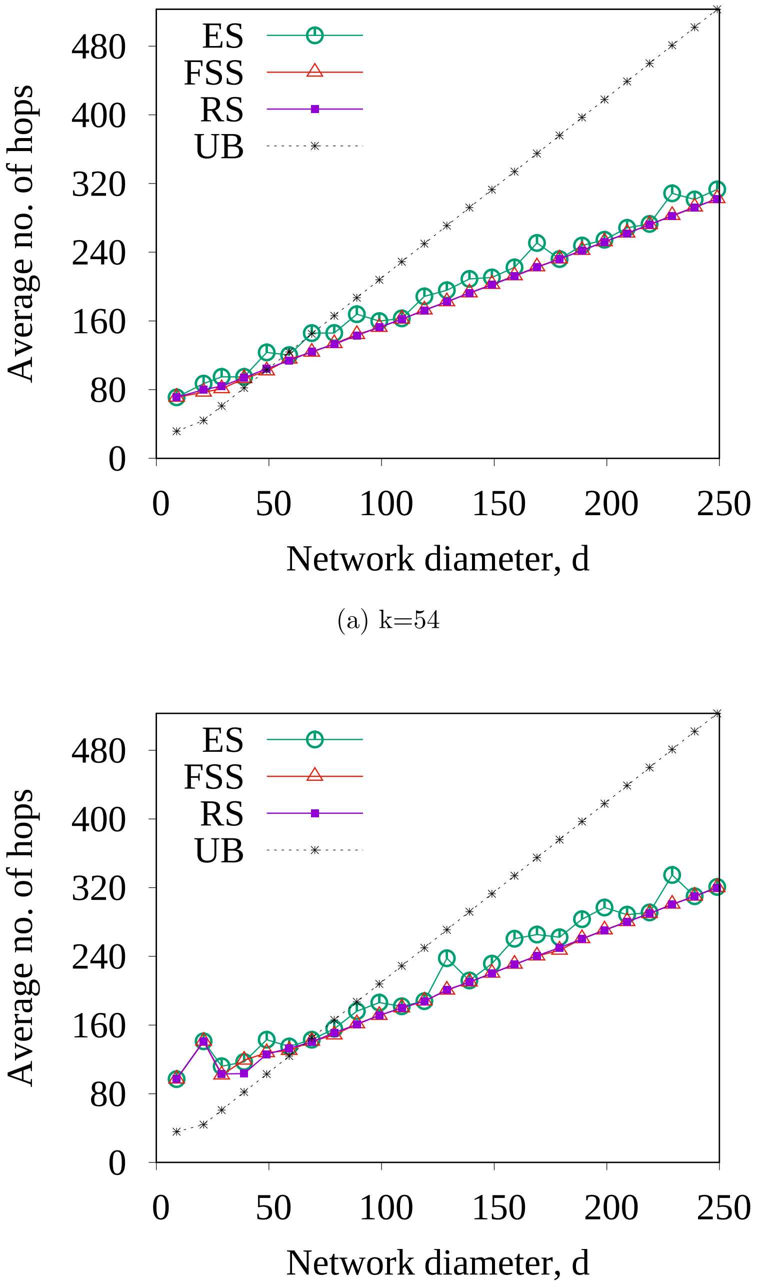
(b) k=72

Figure 6.13: The benefit-cost metric using the worst-case communication-cost ROI_w with varying d .

Path Length

We now compare the path length of the query and response measured in number of hops in the average and worst-case when d is variable for different anonymity-sets construction methods, and several values of k. First of all, we observed that empirical results respected our theoretical analysis in Theorem 10. Specifically, we found that the path length is in $O(d)$ for $k \leq d$. We further estimated the value of the asymptotic constants to be 2.1 for the average-case, and 2.9 for the worst-case path length. That is, the path length in the average and worst-case is less that $2.1d$, and $2.9d$ for $k < d$. We depict our results in Fig. 6.14, and Fig. 6.15 for various values of k, and different anonymity-sets construction methods.

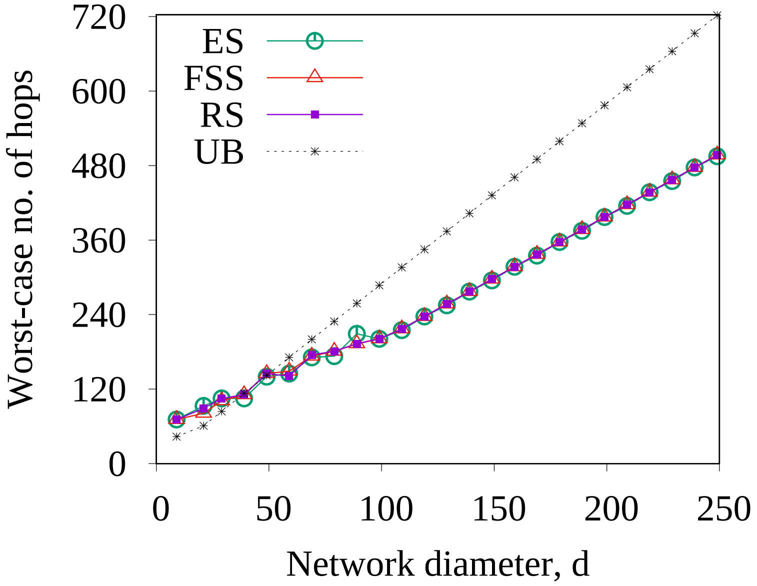
It is found that in the average-case, the generated paths by ES are mostly longer than that by the others, due to the fact that ES distributes the reminder nodes evenly over all the anonymity-sets, thereby subject to the longest path. In the worst-case, on the other hand, the three anonymity-sets construction methods mostly produced identical path lengths as d increased, mostly due to the fixed and relatively small values of k that well mitigates the influence of collection-hops and lets the effect of routing-hops that grows with d rather trivial. Hence, the worst-case path length occurred at an anonymity-set that is far away from the first set where the routing-hops are more. However, the size of such anonymity-set, and eventually the number of collection-hops, is almost the same for different anonymity-sets construction methods. This is because the remainder out of n/k is small, so the anonymity-sets construction methods either change the size of the anonymity-sets other than the first set mildly or leave it untouched. Consequently, the different anonymity-sets construction methods behaved comparably.



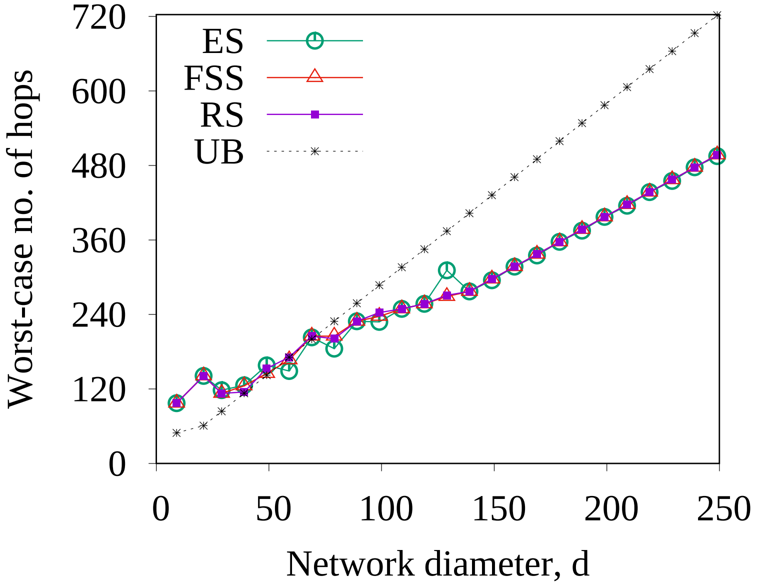
(a) k=54

(b) k=72

Figure 6.14: The average number of hops with varying d .



(a) k=54



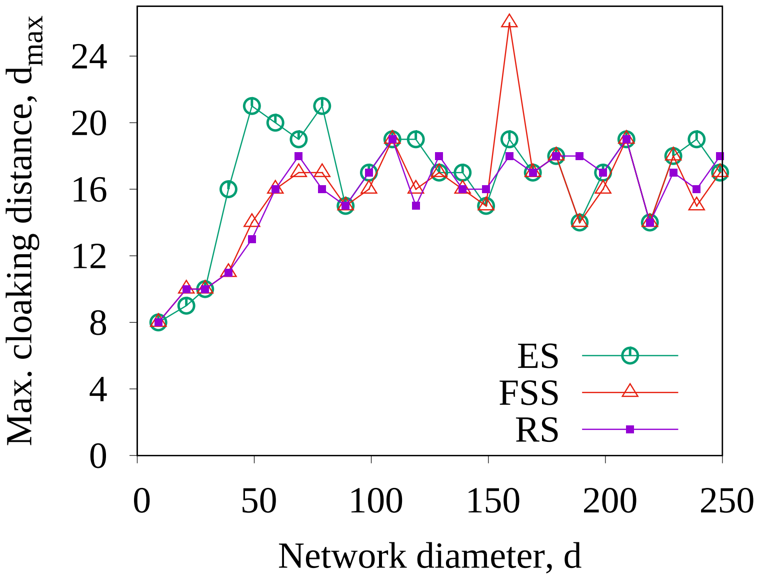
(b) k=72

Figure 6.15: The worst-case number of hops with varying d .

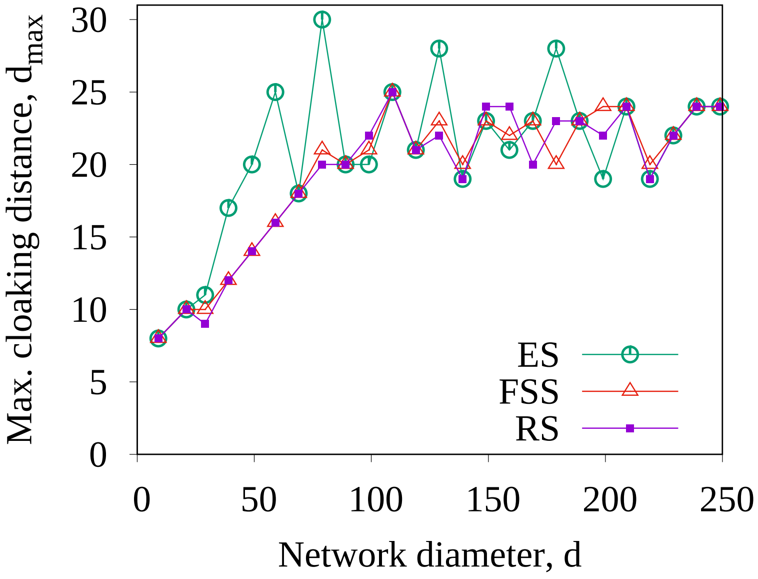
Location Anonymity

Recall that our empirical evaluation in section 6.2.1 confirms the first part of the assertion we make in Theorem 11. To test the second part of the mentioned assertion, we computed the location-anonymity metrics namely, d_{max} and r using Algorithm 4 for large range of d values. We observed that both of the metrics were less than k for relatively large values of d compared to k. This observation respected the upper bound we established in Theorem 11, which is $O(k)$, for $k \in [1, \Theta(d)]$. As argued in the proof of the same theorem, the location-anonymity offered by secure k-anonymous query scheme is limited by the anonymity-set size which is in turn asymptotically bounded from above by the level-of-anonymity k. Our results are depicted in Fig. 6.16, and Fig. 6.17.

By comparing location-anonymity metrics under different anonymity-sets construction methods, we noted similar observations to that of varying level-of anonymity k (see section 6.2.1). That is, ES achieved better location-anonymity compared to FSS whenever there is a remainder out of n/k for the reason mentioned section 6.2.1. We also found that as d is large they all performed similarly. This is due to the fact that as d and n are increased, the number of anonymity-sets $\lfloor n/k \rfloor$ exceeds the remainder out of n/k since we kept k fixed during this set of experiment. This possibly causes ES unable to distribute the remainder equally on all the anonymity-sets, and possibly some of the anonymity-sets would be left untouched at size of $|s| = k$ which is also the smallest size of anonymity-sets for FSS. Recall that we select the anonymity-set with the minimum size to compute d_{max} and r. Thus, FSS and ES behaves similarly. However, they may differ a little due to the variations in the geometric structure of the anonymity-sets that affects the cloaking distances measurements within an anonymity-set.

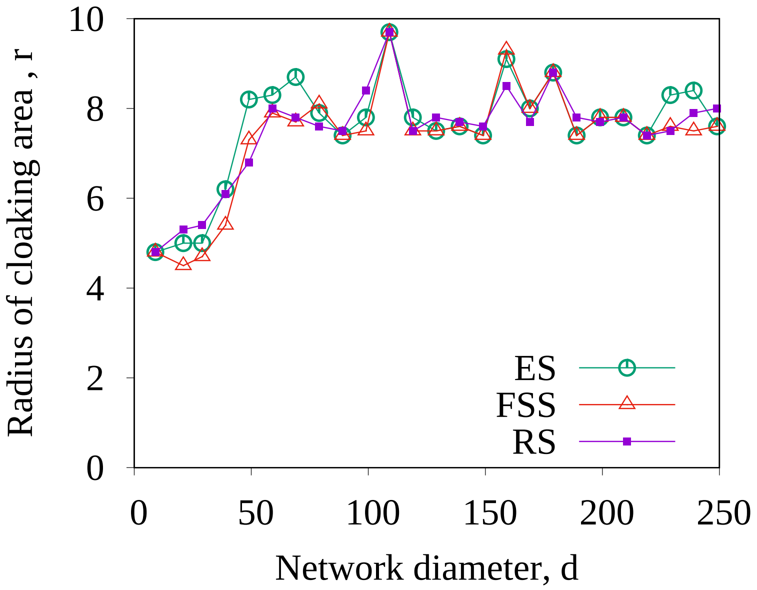


(a) k=54

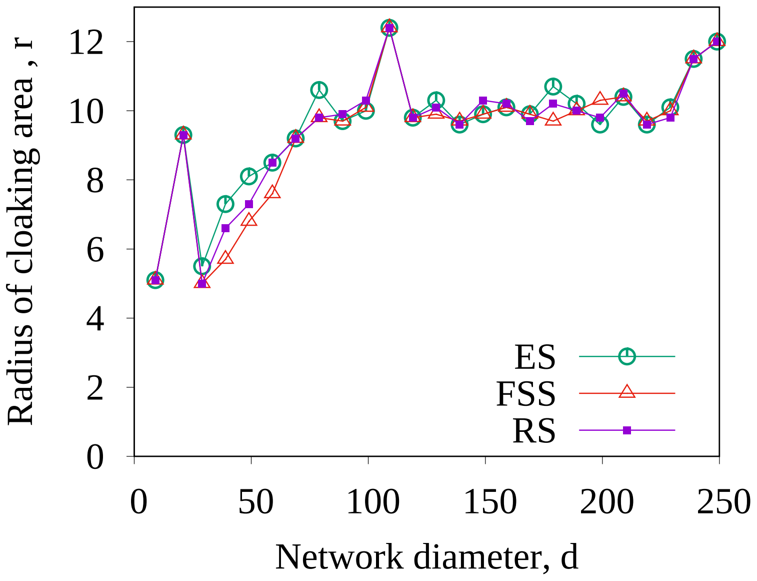


(b) k=72

Figure 6.16: The maximum cloaking distance metric d_{max} with varying network diameter d .



(a) k=54



(b) k=72

Figure 6.17: The radius of equivalent cloaking area metric r with varying network diameter d .

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this work, we studied the average and worst-case analysis of the trade-off between query-anonymity and communication-cost in the context of sensor-cloud based IoT systems. Towards that, we proposed a novel theoretical framework for secure k-anonymous query schemes on the basis of well-known security models namely, ciphertext indistinguishability under chosen plaintext attack (IND-CPA), and information theoretic notion of perfect secrecy. Along the way, we introduced a new notion of secure query k-anonymity that relates itself to the standard concept of anonymity-set in a natural way. This framework is proved to be useful in the design of secure k-anonymous query scheme, analysis of trade-off between query-anonymity and communication-cost, and evaluation of provably secure query-anonymity protocols that are immune to intersection attack, and other types of traffic analysis attacks. We show that constructing equi-probable anonymity sets is not only sufficient for achieving a certain level-of-anonymity, but also the cheapest manner in which to achieve it.

To prove its validity, we presented DAS scheme as a real-world realization of our theoretical framework that achieves secure query k-anonymity in the presence of an eavesdropping adversary who intercepts the anonymity scheme for sufficiently long time. In DAS, we adopt a unique proactive approach that avoids intersection attack by using disjoint anonymity-sets with distinct members. We break DAS down into its essential elements namely, the partition algorithm π, the anonymity transformation T, and the inverse transformation T^{-1}. This proves to be useful in our modular design of DAS for the sensor-cloud-based IoT environment, especially in the implementation of its partition algorithm, query routing algorithm, and querying protocol. The main design goal is to lower the incurred communication-cost while satisfying the desired level-of-anonymity k.

Despite its simplicity, DAS is remarkably useful in the analysis of trade-off between offered query k-anonymity and incurred communication-cost. By offering secure query k-anonymity, DAS precludes all types of intersection attacks, and thus assures a certain level-of-anonymity regardless of the computational power of the adversary. DAS also facilitates the query-anonymity measurement model since its offered level-of-anonymity k is equal to the size of the anonymity-set, i.e. $k = |s|$.

Our theoretical assertions establishes that the network diameter d plays the role of the system-wide inflection point $k \in \Theta(d)$ at which the asymptotic growth of the average and worst-case communication-cost changes from d^2 into k^2 dominance. That is, the asymptotic bounds on the average and worst-case communication-cost is quadratically increased as the offered level-of-anonymity exceeds the network diameter, and they are quadratic in the network diameter for the opposite range.

Based on the proposed querying and routing protocols, extensive analysis was conducted in order to establish bounds on trade-offs between offered query-anonymity and various performance measures such as communication-cost, return-on-investment ROI, path length, and offered location-anonymity. Such comprehensive evaluation helps to unravel key factors affecting the performance of the anonymity scheme. Our experimental results show that the return-on-investment ROI is at its highest values at the point $k = d$, and most of the deduced bounds are functions of the level-of-anonymity k and network diameter d. For instance, both of the location-anonymity metrics d_{max} and r are in $O(d)$ for larger values of k with respect to d, and in $O(k)$ for the opposite range. On the other hand, the query and response path length measured in number of hops is in $O(k)$ in the average-case and worst-case when k surpasses d, and it is in $O(d)$ for the opposite range.

Furthermore, the performance of the different anonymity-sets construction methods was observed and analyzed, which shows that in the average-case scenario, First-Set-Spread outperforms when k is relatively small compared to d, while Equal-Spread is favored when k is relatively large compared to d. In the worst-case scenario, First-Set-Spread fails behind Equal-Spread and this is more visible at large values of k.

7.2 Future Work

We endeavor to make our evaluation results more comprehensive by testing our ideas in various simulation settings of routing algorithms, data collection strategies, network topologies and other system parameters. In this thesis, we have considered source routing and piggyback data collection in a square grid topology. It will be interesting to explore the behavior of the secure k-anonymous scheme under a variety of routing algorithms such as dynamic routing, data collections that do not use piggyback strategy, and different network environments.

Another streak of research is to design different partition algorithms that make use of the security properties of secure k-anonymous query scheme. In this work so far, we considered disjoint anonymity-sets whose members are adjacent to implement our partition algorithm π. Investigation of other approaches that adopt nonadjacent disjoint anonymity-sets, and comparison with our approach is a fruitful area of research to explore. Such approaches are deemed to be efficient in terms of the achieved location-anonymity since the nodes in each anonymity-set are remote from each other. Also, it will be then required to develop suitable static or adaptive techniques to spread the members of each anonymity-set all over the network in order to achieve a certain level of location-anonymity. Our location-anonymity metrics namely, radius of equivalent cloaking area r, and maximum cloaking distance d_{max} can be utilized as crucial tools in the analysis and measurement of offered location-anonymity.

When the nonadjacent disjoint anonymity-sets of DAS are put into effect, it will then be interesting to seek rigorous answers to the following research questions using formal methods:

- To achieve a secure query k-anonymity of level k, what is the sufficient communication-cost on average and in the worst-case?
- To achieve a secure query k-anonymity of level k, what is the necessary communication-cost on average and in the worst-case?
- What is the impact of spreading members of disjoint anonymity-sets throughout the network on the performance of secure query k-anonymity scheme?

In the formulation of the cost-anonymity trade-off problem, we primarily focused on global passive eavesdropping adversary, specifically *honest-but-curious* model. As another direction of future work, we may consider other models of attacker such as active adversary. Such adversary may be able to manipulate the query and response messages, or have partial of full control on the physical and virtual environment of sensor-cloud-based IoT System. It will be interesting to ask how we can develop the right strategies to detect malicious traffic or nodes, prevent such active attacks, and recover from their consequences.

In this work, we have avoided intersection and statistical disclosure attack by using disjoint anonymity-sets to allow accurate and precise measurement of the level-of-anonymity since it is equal to the lowest size of anonymity-sets. However, it will be interesting to study the anonymity-performance trade-off under vulnerable environment in which anonymity-sets are allowed to intersect with each other. Then, it will be challenging to design an adaptive partition algorithm to keep the level-of-anonymity within a predefined tolerance as a design parameter.

It can be argued that communication-based anonymity schemes are not the only ones that can implement trust-none model in practice. specifically, homomorphic encryption technique, which allows computations to be carried out directly on the ciphertext, is another mechanism that does not necessitate trusting other components in the system. Consequently, it is also of interest to examine the trade-off problem using this powerful technique.

Another avenue for further research is to develop a fault tolerance mechanism to allow graceful degradation of system performance in case of failure. This requires to design error detection techniques to detect faults, and change or update the routes to queried anonymity-set, and partition algorithm to maintain the same desired level-of-anonymity. Similar adaptive partition and routing algorithms may be feasible to accommodate dynamic network environment in which nodes are allowed to join and leave during the lifetime of the k-anonymous query scheme. The main design goal of these adaptive algorithms is to repartition the network into disjoint anonymity-sets of size in $[k, 2k-1]$.

We also see applications of our theoretical framework, design and implementation of secure query k-anonymity schemes, and our trade-off analysis as a promising future work. For example, we seek to study the performance-anonymity trade-offs problem in various environments such as for big data, and cloud computing. It will then be required to revisit our theoretical framework and our analysis of trade-offs to establish a suite of average and worst-case bounds on them.

References

[1] L. Atzori, A. Iera, and G. Morabito, “The internet of things: A survey,” *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[2] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, “Sensing as a service model for smart cities supported by internet of things,” *Transactions on Emerging Telecommunications Technologies*, vol. 25, no. 1, pp. 81–93, 2014.

[3] M. Yuriyama and T. Kushida, “Sensor-cloud infrastructure-physical sensor management with virtualized sensors on cloud computing,” in *Network-Based Information Systems (NBiS), 2010 13th International Conference on*, pp. 1–8, IEEE, 2010.

[4] X. Sheng, J. Tang, X. Xiao, and G. Xue, “Sensing as a service: Challenges, solutions and future directions,” *IEEE Sensors journal*, vol. 13, no. 10, pp. 3733–3741, 2013.

[5] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of things (iot): A vision, architectural elements, and future directions,” *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[6] S. S. Mathew, Y. Atif, Q. Z. Sheng, and Z. Maamar, “The web of things-challenges and enabling technologies,” in *Internet of things and inter-cooperative computational technologies for collective intelligence*, pp. 1–23, Springer, 2013.

[7] L. von Ahn, A. Bortz, and N. J. Hopper, "K-anonymous message transmission," in *Proceedings of the 10th ACM Conference on Computer and Communications Security*, CCS '03, (New York, NY, USA), pp. 122–130, ACM, 2003.

[8] B. Carbunar, Y. Yu, W. Shi, M. Pearce, and V. Vasudevan, "Query privacy in wireless sensor networks," *ACM Trans. Sen. Netw.*, vol. 6, pp. 14:1–14:34, Mar. 2010.

[9] E. De Cristofaro, X. Ding, and G. Tsudik, "Privacy-preserving querying in sensor networks," in *Computer Communications and Networks, 2009. ICCCN 2009. Proceedings of 18th Internatonal Conference on*, pp. 1–6, Aug 2009.

[10] A. A. Abbasi and M. Younis, "A survey on clustering algorithms for wireless sensor networks," *Computer communications*, vol. 30, no. 14, pp. 2826–2841, 2007.

[11] K. Akkaya and M. Younis, "A survey on routing protocols for wireless sensor networks," *Ad hoc networks*, vol. 3, no. 3, pp. 325–349, 2005.

[12] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *System sciences, 2000. Proceedings of the 33rd annual Hawaii international conference on*, pp. 10–pp, IEEE, 2000.

[13] O. Younis and S. Fahmy, "Heed: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks," *Mobile Computing, IEEE Transactions on*, vol. 3, no. 4, pp. 366–379, 2004.

[14] K. Hayawi, A. Mortezaei, and M. V. Tripunitara, "The limits of the trade-off between query-anonymity and communication-cost in wireless sensor networks," in *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, CODASPY '15, (New York, NY, USA), pp. 337–348, ACM, 2015.

[15] O. Goldreich, *Foundations of cryptography: volume 2, basic applications*. Cambridge university press, 2004.

[16] C. E. Shannon, "Communication theory of secrecy systems*," *Bell system technical journal*, vol. 28, no. 4, pp. 656–715, 1949.

[17] K. Hayawi, P.-H. Ho, S. Mathew, and L. Peng, "Securing the internet of things: a worst-case analysis of trade-off between query-anonymity and communication-cost," in *31st International Conference on Advanced Information Networking and Applications (AINA'17)*, IEEE, 2017.

[18] K. Hayawi, P.-H. Ho, and S. Mathew, "Cost analysis of query-anonymity on the internet of things," *IEEE Transactions on Mobile Computing*, 2017, submitted for publication.

[19] K. Hayawi, P.-H. Ho, S. Mathew, and L. Peng, "Secure k-anonymity query scheme on the internet of things: Design and performance analysis," *IEEE/ACM Transactions on Networking*, submitted for publication.

[20] D. L. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communications of the ACM*, vol. 24, no. 2, pp. 84–90, 1981.

[21] D. Chaum, "The dining cryptographers problem: Unconditional sender and recipient untraceability," *Journal of cryptology*, vol. 1, no. 1, pp. 65–75, 1988.

[22] D. Goldschlag, M. Reed, and P. Syverson, "Onion routing," *Communications of the ACM*, vol. 42, no. 2, pp. 39–41, 1999.

[23] N. Mathewson, P. Syverson, and R. Dingledine, "Tor: the second-generation onion router," in *Proc. USENIX Security Symp*, 2004.

[24] M. Edman and B. Yener, "On anonymity in an electronic society: A survey of anonymous communication systems," *ACM Computing Surveys (CSUR)*, vol. 42, no. 1, p. 5, 2009.

[25] R. Dingledine, V. Shmatikov, and P. Syverson, "Synchronous batching: From cascades to free routes," in *Privacy Enhancing Technologies*, pp. 186–206, Springer, 2005.

[26] A. Serjantov, R. Dingledine, and P. Syverson, "From a trickle to a flood: Active attacks on several mix types," in *Information Hiding*, pp. 36–52, Springer, 2003.

[27] C. Diaz and A. Serjantov, "Generalising mixes," in *Privacy Enhancing Technologies*, pp. 18–31, Springer, 2003.

[28] D. Kesdogan, J. Egner, and R. Büschkes, "Stop-and-go-mixes providing probabilistic anonymity in an open system," in *Information Hiding*, pp. 83–98, Springer, 1998.

[29] L. OConnor, "On blending attacks for mixes with memory," in *Information Hiding*, pp. 39–52, Springer, 2005.

[30] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Communications of the ACM*, vol. 21, no. 2, pp. 120–126, 1978.

[31] B. Pfitzmann and A. Pfitzmann, "How to break the direct rsa-implementation of mixes," in *Advances in CryptologyEUROCRYPT89*, pp. 373–381, Springer, 1990.

[32] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," in *Advances in Cryptology*, pp. 10–18, Springer, 1985.

[33] G. Danezis and C. Diaz, "A survey of anonymous communication channels," *Computer Communications*, vol. 33, 2008.

[34] T. Dierks, "The transport layer security (tls) protocol version 1.2," 2008.

[35] M. K. Wright, M. Adler, B. N. Levine, and C. Shields, "Passive-logging attacks against anonymous communications systems," *ACM Transactions on Information and System Security (TISSEC)*, vol. 11, no. 2, p. 3, 2008.

[36] O. Berthold, A. Pfitzmann, and R. Standtke, "The disadvantages of free mix routes and how to overcome them," in *Designing Privacy Enhancing Technologies*, pp. 30–45, Springer, 2001.

[37] D. Kedogan, D. Agrawal, and S. Penz, "Limits of anonymity in open environments," in *International Workshop on Information Hiding*, pp. 53–69, Springer, 2002.

[38] N. Mathewson and R. Dingledine, "Practical traffic analysis: Extending and resisting statistical disclosure," in *Privacy Enhancing Technologies*, pp. 17–34, Springer, 2005.

[39] M. K. Wright, M. Adler, B. N. Levine, and C. Shields, "The predecessor attack: An analysis of a threat to anonymous communications systems," *ACM Transactions on Information and System Security (TISSEC)*, vol. 7, no. 4, pp. 489–522, 2004.

[40] V. Shmatikov, "Probabilistic analysis of anonymity," in *Computer Security Foundations Workshop, 2002. Proceedings. 15th IEEE*, pp. 119–128, IEEE, 2002.

[41] J.-F. Raymond, "Traffic analysis: Protocols, attacks, design issues, and open problems," in *Designing Privacy Enhancing Technologies*, pp. 10–29, Springer, 2001.

[42] G. Danezis, C. Diaz, and C. Troncoso, "Two-sided statistical disclosure attack," in *Privacy Enhancing Technologies*, pp. 30–44, Springer, 2007.

[43] B. N. Levine, M. K. Reiter, C. Wang, and M. Wright, "Timing attacks in low-latency mix systems," in *International Conference on Financial Cryptography*, pp. 251–265, Springer, 2004.

[44] L. Overlier and P. Syverson, "Locating hidden servers," in *Security and Privacy, 2006 IEEE Symposium on*, pp. 15–pp, IEEE, 2006.

[45] L. Øverlier and P. Syverson, "Valet services: Improving hidden servers with a personal touch," in *Privacy Enhancing Technologies*, pp. 223–244, Springer, 2006.

[46] U. Möller, L. Cottrell, P. Palfrader, and L. Sassaman, "Mixmaster protocolversion 2," *Draft, July*, 2003.

[47] G. Danezis, R. Dingledine, and N. Mathewson, "Mixminion: Design of a type iii anonymous remailer protocol," in *Security and Privacy, 2003. Proceedings. 2003 Symposium on*, pp. 2–15, IEEE, 2003.

[48] M. Waidner, B. Pfitzmann, *et al.*, "The dining cryptographers in the disco: Unconditional sender and recipient untraceability with computationally secure serviceability," *J.-J. Quisquater and J. Vandewalle, editors, Advances in CryptologyEUROCRYPT*, vol. 89, p. 690, 1989.

[49] S. Dolev and R. Ostrobsky, "Xor-trees for efficient anonymous multicast and reception," *ACM Transactions on Information and System Security (TISSEC)*, vol. 3, no. 2, pp. 63–84, 2000.

[50] M. K. Reiter and A. D. Rubin, "Crowds: Anonymity for web transactions," *ACM Transactions on Information and System Security (TISSEC)*, vol. 1, no. 1, pp. 66–92, 1998.

[51] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[52] S. Goel, M. Robson, M. Polte, and E. Sirer, "Herbivore: A scalable and efficient protocol for anonymous communication," tech. rep., Cornell University, 2003.

[53] E. G. Sirer, M. Polte, M. Robson, E. Gün, S. Milo, and P. M. Robson, "Cliquenet: A self-organizing, scalable, peer-to-peer anonymous communication substrate," 2001.

[54] E. De Cristofaro, X. Ding, and G. Tsudik, "Privacy-preserving querying in sensor networks," in *Proceedings of 18th International Conference on Computer Communications and Networks (ICCCN 2009)*, pp. 1–6, aug. 2009.

[55] N. Li, N. Zhang, S. K. Das, and B. Thuraisingham, "Privacy preservation in wireless sensor networks: A state-of-the-art survey," *Ad Hoc Networks*, vol. 7, no. 8, pp. 1501–1514, 2009.

[56] A. Pfitzmann and M. Hansen, "A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management," 2010.

[57] C. Diaz, S. Seys, J. Claessens, and B. Preneel, "Towards measuring anonymity," in *Privacy Enhancing Technologies*, pp. 54–68, Springer, 2003.

[58] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *Privacy Enhancing Technologies*, pp. 41–53, Springer, 2003.

[59] A. Pfitzmann and M. Hansen, “A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management,” 2010.

[60] A. Serjantov and G. Danezis, “Towards an information theoretic metric for anonymity,” in *Proceedings of the 2nd international conference on Privacy enhancing technologies*, PET’02, (Berlin, Heidelberg), pp. 41–53, Springer-Verlag, 2003.

[61] C. Díaz, S. Seys, J. Claessens, and B. Preneel, “Towards measuring anonymity,” in *Proceedings of the 2nd international conference on Privacy enhancing technologies*, PET’02, (Berlin, Heidelberg), pp. 54–68, Springer-Verlag, 2003.

[62] D. Kedogan, D. Agrawal, and S. Penz, “Limits of anonymity in open environments,” in *Information Hiding*, pp. 53–69, Springer, 2003.

[63] J.-F. Raymond, “Traffic analysis: Protocols, attacks, design issues, and open problems,” in *Designing Privacy Enhancing Technologies*, pp. 10–29, Springer, 2001.

[64] D. Agrawal and D. Kesdogan, “Measuring anonymity: The disclosure attack,” *IEEE Security & privacy*, vol. 1, no. 6, pp. 27–34, 2003.

[65] M. Wright, M. Adler, B. N. Levine, and C. Shields, “Defending anonymous communications against passive logging attacks,” in *Security and Privacy, 2003. Proceedings. 2003 Symposium on*, pp. 28–41, IEEE, 2003.

[66] G. Tóth, Z. Hornák, and F. Vajda, “Measuring anonymity revisited,” in *Proceedings of the Ninth Nordic Workshop on Secure IT Systems*, pp. 85–90, Espoo, Finland, 2004.

[67] A. Pfitzmann and M. Köhntopp, “Anonymity, unobservability, and pseudonymitya proposal for terminology,” in *Designing privacy enhancing technologies*, pp. 1–9, Springer, 2001.

[68] C. E. Shannon, “A mathematical theory of communication,” *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[69] S. Clauß and S. Schiffner, “Structuring anonymity metrics,” in *Proceedings of the second ACM workshop on Digital identity management*, pp. 55–62, ACM, 2006.

[70] Y. Deng, J. Pang, and P. Wu, “Measuring anonymity with relative entropy,” in *Formal Aspects in Security and Trust*, pp. 65–79, Springer, 2007.

[71] K. Chatzikokolakis, C. Palamidessi, and P. Panangaden, “Anonymity protocols as noisy channels,” in *Trustworthy Global Computing*, pp. 281–300, Springer, 2007.

[72] I. S. Moskowitz, R. E. Newman, D. P. Crepeau, and A. R. Miller, “Covert channels and anonymizing networks,” in *Proceedings of the 2003 ACM workshop on Privacy in the electronic society*, pp. 79–88, ACM, 2003.

[73] G. Danezis and A. Serjantov, “Statistical disclosure or intersection attacks on anonymity systems,” in *Information Hiding*, pp. 293–308, Springer, 2005.

[74] A. Hevia and D. Micciancio, “An indistinguishability-based characterization of anonymous channels,” in *Privacy Enhancing Technologies*, pp. 24–43, Springer, 2008.

[75] J.-M. Bohli and A. Pashalidis, “Relations among privacy notions,” in *Financial Cryptography and Data Security*, pp. 362–380, Springer, 2009.

[76] N. Gelernter and A. Herzberg, “On the limits of provable anonymity,” in *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, pp. 225–236, ACM, 2013.

[77] P. Golle and A. Juels, “Dining cryptographers revisited,” in *Advances in Cryptology-Eurocrypt 2004*, pp. 456–473, Springer, 2004.

[78] A. Beimel and S. Dolev, “Buses for anonymous message delivery,” *Journal of Cryptology*, vol. 16, no. 1, pp. 25–39, 2003.

[79] S. Goldwasser and S. Micali, “Probabilistic encryption,” *Journal of computer and system sciences*, vol. 28, no. 2, pp. 270–299, 1984.

[80] M. Bellare, A. Desai, D. Pointcheval, and P. Rogaway, “Relations among notions of security for public-key encryption schemes,” in *Advances in CryptologyCRYPTO'98*, pp. 26–45, Springer, 1998.

[81] J. Katz and Y. Lindell, *Introduction to modern cryptography.* CRC Press, 2014.

[82] S. Goldwasser and S. Micali, “Probabilistic encryption,” *Journal of Computer and System Sciences*, vol. 28, no. 2, pp. 270 – 299, 1984.

[83] C. P. Pfleeger and S. L. Pfleeger, *Security in computing.* Prentice Hall Professional Technical Reference, 2002.

[84] C. E. Shannon, “A mathematical theory of communication,” *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, pp. 3–55, Jan. 2001.

[85] R. Anderson, *Security engineering.* John Wiley & Sons, 2008.

[86] D. R. Stinson, *Cryptography: theory and practice.* CRC press, 2005.

[87] D. Kesdogan and L. Pimenidis, “The hitting set attack on anonymity protocols,” in *Information Hiding*, pp. 326–339, Springer, 2005.

[88] E. F. Krause, *Taxicab geometry: An adventure in non-Euclidean geometry.* Courier Corporation, 2012.

[89] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms (3. ed.).* MIT Press, 2009.

[90] K. Bogart, S. Drysdale, and C. Stein, “Discrete math for computer science students,” 2004.

[91] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms.* MIT press, 2009.

[92] S. M. Ross, *Introduction to probability models.* Academic press, 2014.

[93] M. Gruteser and D. Grunwald, “Anonymous usage of location-based services through spatial and temporal cloaking,” in *Proceedings of the 1st international conference on Mobile systems, applications and services*, pp. 31–42, ACM, 2003.