

# Defining Privacy Levels for IP Address Anonymization

**Wongyos Keardsri**<sup>1, C1</sup> **Yunyong Teng-amnuay**<sup>1, C2</sup> and **Passakon Prathombutr**<sup>2, C3</sup>

<sup>1</sup> *Information System Engineering Laboratory (ISEL), Center of Excellence in Software Engineering  
Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University  
Phayathai Road, Pathumwan, Bangkok 10330, Thailand*

<sup>2</sup> *National Electronics and Computer Technology Center (NECTEC)  
National Science and Technology Development Agency (NSTDA), Ministry of Science and Technology  
Thailand Science Park, Phaholyothin Road, Klong Luang, Pathumthani 12120, Thailand*

**E-mail:** <sup>C1</sup>g49wkr@cp.eng.chula.ac.th, <sup>C2</sup>Yunyong.T@Chula.ac.th, <sup>C3</sup>prathom@nectec.or.th

## ABSTRACT

Nowadays, an IP address anonymization is an important technique for network analysis and Internet research. The method of anonymization is the changing of original IP address to anonymized IP address. This can prevent sensitive information of users from disclosure. However, most current anonymization techniques are unsuitable for network analysis functions. They anonymize all 32 bits of IP address unnecessarily. In this paper, we propose 5 privacy levels that anonymize a part of IP address in a different scheme. We apply these privacy levels to prefix-preserving IP address anonymization. Our anonymization scheme benefits any organizations in exchanging network data, and also appropriate for packet tracers and sniffers.

**Keywords:** IP address anonymization, privacy, privacy levels, sensitive information, network analysis, Internet research, packet tracer, packet sniffer

## INTRODUCTION

Nowadays, packet tracers and sniffers are a crucial tool for network analysis and Internet research such as traffic analysis, system diagnosis, network performance evaluation, and more generally network analysis functions, to analyze and evaluate the condition of network system. The packet data from the traces which contain the source and destination IP addresses can link to users who are in the network. To prevent user privacy which may be inferred from the trace, the IP address must be removed or closed by using an anonymization technique. The IP address anonymization is the replacing of original IP address to anonymized IP address to keep the private information of users in network and to prevent suitable a disclosure and violation of user privacy. The well-known anonymization techniques are TCPdpriv [1], Crypto-PAn [2], Multiple Access Level [3], and TSA [4]; however, they are unsuitable for network analysis functions. Because they do not consider the appropriate bits or parts of IP address to anonymize and also anonymize all 32 bits of IP address unnecessarily. In fact, the anonymization depend on the packet data which need to analyze and parts of IP address which need to see.

In this paper, we anonymize the necessary bits or parts of IP address according to different privacy levels and views. We propose 5 privacy levels for anonymization scheme. The first level is non-anonymization; all 32 bits of IP address are not anonymized. The second level is n-left anonymization; only n bits of IP address from network part are anonymized. The third level is n-right anonymization; only n bits of IP address from host part are anonymized. The fourth level is full anonymization; all 32 bits of IP address, which consist of host and network parts, are anonymized. The last level is randomly full anonymization; all 32 bits of IP address are randomized before being anonymized. We apply 5 privacy levels to prefix-preserving IP address anonymization, the technique which can preserve network relationship among the same network domain from original IP addresses. Our anonymization scheme is applicable to an administrator who analyzes packet data. The scheme benefits any organizations in exchanging network data, and also appropriate for packet tracers and sniffers.

## THEORY AND RELATED WORKS

### IP Address Structure

An IP address is a data structure which uniquely identifies a node within the network. Each IP address consists of two parts. First is a network part; it identifies a network domain and a group of organization. Second is a host part; it identifies a host number within a local network of each organization. In the IPv4 [5], network address is identified as a bit-wise logical AND of the 32-bit IPv4 address with 32-bit subnet mask or netmask address. The netmask has a “1” bit for each bit that is a part of the network number, and a “0” bit for each bit which is a part of the host number. All systems within the same network domain share the same subnet mask address. When each IP address is calculated with subnet mask address, it can be separated into network and host parts. These above-mentioned details show in Figure 1.

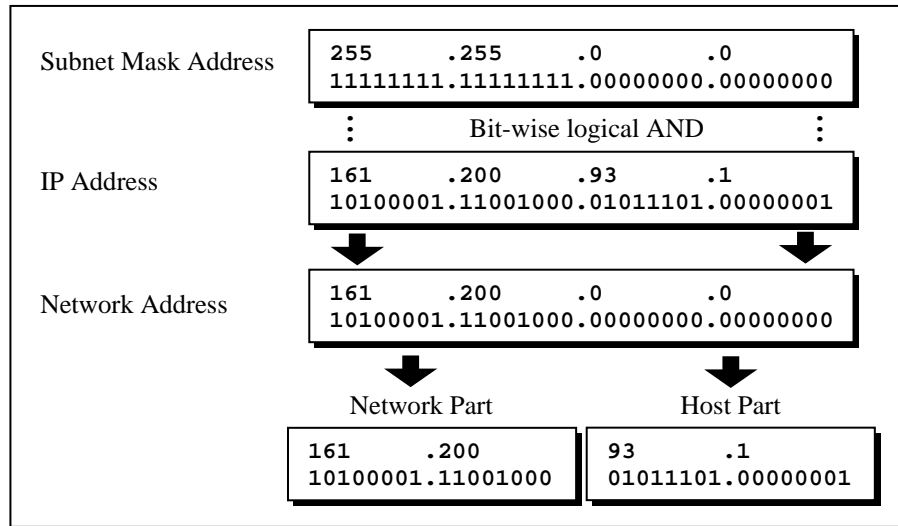


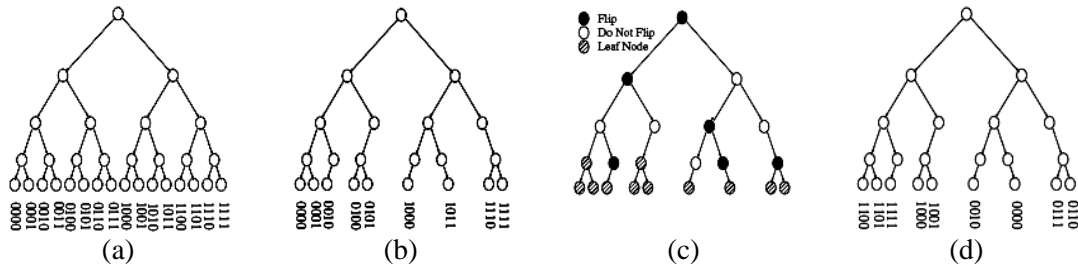
Figure 1. IP address structure

### IP Address Anonymization

An IP address anonymization was first proposed in 1990s. The basic approach is one-to-one mapping [1], which maps each IP address to random 32 bits of IP address. The examples of this approach are Blowfish [6], MD5 [6, 7], and Hashing function. One-to-one mapping can only preserve privacy requirement, however, the results from anonymization cannot preserve the prefix relationships among original IP addresses. To apply anonymization scheme to be prefix preserving, it should anonymize two of original IP addresses to share a k-bit prefix, and their anonymized mappings will also share a k-bit prefix. First approach to such prefix preserving anonymization is a TCPdpriv (tcpdump privacy). It was developed by Greg Minshall [1]. The TCPdpriv was implemented to anonymize IP addresses with different security levels. In level 0, it maps different addresses to integers (counting from 1). In level 1, it maps the upper and lower 16 bits of IP address, separately, to integers (counting from 1); the upper and lower maps are independent. In level 2, it maps each byte of the address separately; each byte map is independent. Level 50 is prefix-preserving anonymization scheme. The TCPdpriv can be viewed as a table-based approach, it stores a set of raw IP address, and anonymized IP address binding pairs of IP addresses to maintain the consistency of the anonymization within one trace. The binding is generated randomly when a new address is anonymized. Therefore, it may produce an inconsistency of anonymized address (i.e., same original prefix is mapped into different anonymized prefixes) when the same IP address is used more than one trace.

To improve TCPdpriv scheme, Jun Xu [2] et al proposed a new prefix-preserving anonymization by using cryptographic concepts. They called Crypto-PAn. It is the well-known deterministic prefix preserving anonymization which maps raw addresses to anonymized addresses with the same key, so it can anonymize IP address consistently. The Crypto-PAn algorithm is based on the canonical form theorem. It represents the distinct IPv4 addresses with a *complete binary tree* of height 32 as show in Figure 2(a). The set of distinct original addresses

can be represented by *original address tree* as show in Figure 2(b). It uses *anonymization function* in Figure 2(c) to anonymize original IP address. This function can be generated by specifying a binary variable for each non-leaf node (including the root node) of the original address tree. This variable specifies whether the function "flips" this bit or not. The result from anonymization function is the *anonymized address tree* as show in Figure 2(d).



**Figure 2.** Address trees and anonymization function: (a) complete binary tree (using 4-bit addresses for simplicity); (b) original address tree; (c) anonymization function; (d) anonymized address tree.

The Crypto-PAn uses cryptographic key in 32-bit steps (e.g., 32-bit, 64-bit, 128-bit, etc.) to anonymization function. The same address which appears in two different traces will be mapped to the same anonymized address if the same key is used. Therefore, this scheme is consistent in prefix-preserving anonymization. However, Crypto-PAn is low efficiency for a heavy packet tracer on real time high speed network because of time which generates address trees is long. It needs 32 rounds of encryption, thus makes it unsuitable for real time anonymization.

The prefix-preserving anonymization is an interesting method to many researchers. Qianli Zhang [3] and et al are one which proposed the anonymization scheme by using multiple access levels of the traces. This scheme is made more secure by different keys, but it may be complex to access data if the levels and keys increase. Therefore, they are complicated to recover and unsuitable for real time anonymization.

In the previous problem, Ramaswamy Ramaswamy [4] and et al proposed the top-hash subtree-replicated anonymization (TSA), the high-speed prefix-preserving anonymization, by using pre-computation, replicated subtrees, and top hashing, to improve the computation time of anonymization algorithm. Moreover, Qianli Zhang [7] and et al also proposed fast prefix-preserving anonymization by using bit string algorithm to improve anonymization performance.

## PRIVACY LEVELS

Previous proposed techniques always anonymize all 32 bits of IP address. When we studied and surveyed in the network analysis processes, we found that some processes do not require anonymization; some processes require anonymizing only some parts of IP address. Therefore, we can consider the appropriate bit numbers and appropriate parts of IP address to anonymization. Consequently, we propose the privacy levels to anonymization scheme. We define these privacy levels into 5 levels as follows.

### *Non-Anonymization*

*Non-anonymization* is the first level which all 32 bits of original IP address are not anonymized. This level is used to anonymize IP address in packet data which needs to analyze the network summary results such as bandwidth usage, network service summary, and capacity planning. These analysis processes are not related to users or specific network, so it is unnecessary to anonymize the IP address. The anonymized IP address by using non-anonymization level is shown in Figure 3(a).

### *n-Left Anonymization*

*n-Left anonymization* is the second level which only n bits of IP address from network part are anonymized. This level is used to anonymize IP address in packet data which needs to

analyze the results that specify the network part during analysis process such as comparing network resource usage, and comparing network service statistics. The anonymized IP address by using n-left anonymization level is shown in Figure 3(b).

***n-Right Anonymization***

*n-Right anonymization* is the third level which only n bits of IP address from host part are anonymized. This level is used to anonymize IP address in packet data which needs to analyze the results that specify the host part during analysis process such as CPU usage, memory usage, and device services summary. The anonymized IP address by using n-right anonymization level is shown in Figure 3(c).

***Full Anonymization***

*Full anonymization* is the fourth level which all 32 bits of IP address, which consist of the host network parts, are anonymized. This level is used to anonymize IP address in packet data which needs to analyze the results that specify both of network host parts during analysis process such as user behavior analysis, intrusion detection, log analysis, and social network analysis. The anonymized IP address by using full anonymization level is shown in Figure 3(d).

***Randomly Full Anonymization***

*Randomly full anonymization* is the fifth level which all 32 bits of IP address are randomized by random algorithm. This has been done before they are anonymized and be consistent by table lookup. This level is used to anonymize IP address in packet data which needs to analyze the results that do not require prefix-preserving and display the results to the public such as list of device services, web application report, and network map. The anonymized IP address by using randomly full anonymization level is shown in Figure 3(e).

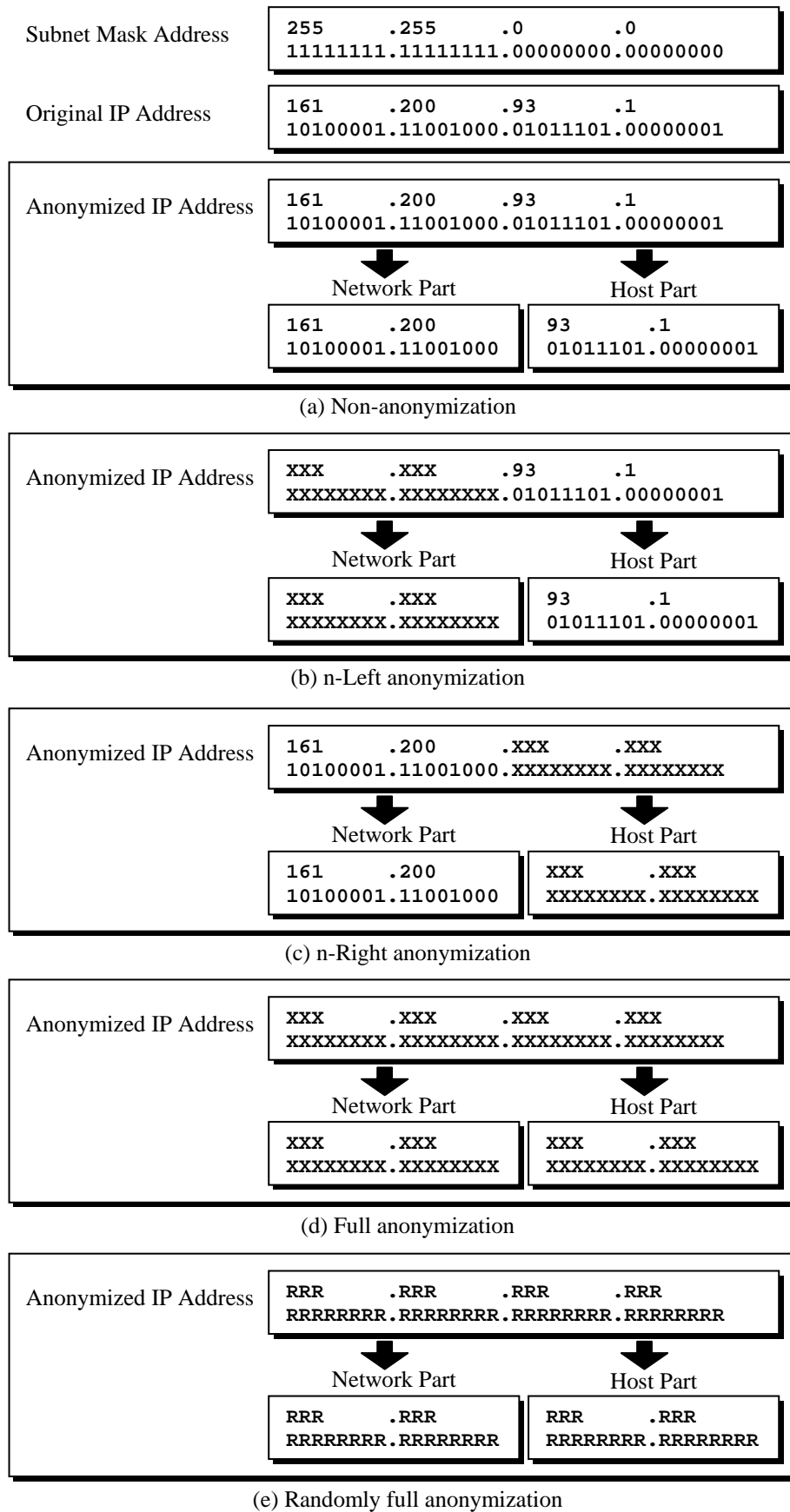
**ANONYMIZATION SCHEME**

We apply the privacy levels to prefix-preserving anonymization, specifically to Crypto-PAn. The Crypto-PAn uses the trees to anonymize all 32 bits of IP address. In our scheme, we implement the anonymization algorithm to anonymize n bits of IP address by applying Crypto-PAn algorithm for n-left, n-right and full anonymization, and using the random algorithm for randomly full anonymization. This algorithm shows in Table 1.

**Table 1.** Anonymization algorithm based on privacy levels

1	<i>originalIP</i> $\leftarrow$ <i>input IP address</i>
2	<i>subnet</i> $\leftarrow$ <i>input subnet mask address</i>
3	<i>netPart</i> $\leftarrow$ <i>binary(originalIP) AND binary(subnet)</i>
4	<i>hostPart</i> $\leftarrow$ <i>substring(netPart.length, 32)</i>
5	<i>level</i> $\leftarrow$ <i>input privacy level</i>
6	<i>if (level is 1)</i>
7	<i>anonymizedIP</i> $\leftarrow$ <i>originalIP</i>
8	<i>else if (level is 2)</i>
9	<i>anonymizedIP</i> $\leftarrow$ <i>cryptopan(netPart)</i>
10	<i>else if (level is 3)</i>
11	<i>anonymizedIP</i> $\leftarrow$ <i>cryptopan(hostPart)</i>
12	<i>else if (level is 4)</i>
13	<i>anonymizedIP</i> $\leftarrow$ <i>cryptopan(originalIP)</i>
14	<i>else if (level is 5)</i>
15	<i>anonymizedIP</i> $\leftarrow$ <i>random(originalIP)</i>
16	<i>else //undefined level</i>
17	<i>anonymizedIP</i> $\leftarrow$ <i>originalIP</i>

This anonymization scheme depends on the packet data which is used to analyze, and is related to a network part, a host part, or unrelated. It can be applied to either batch or real-time anonymization.



**Figure 3.** Privacy levels; (a) non-anonymization; (b) n-left anonymization; (c) n-right anonymization; (d) full anonymization; (e) randomly full anonymization

## RESULTS AND DISCUSSION

The results from our anonymization scheme based on privacy levels are different from the level which is used. Table 2 is an example of possible results from 5 levels.

Given the data which are used in anonymization process are as follows.

1. Network address 161.200.93.0
2. Subnet mask 255.255.255.0

The network in binary format is 10100001110010000101110100000000, and subnet mask is 11111111111111111111111110000000. Therefore, the network part is the first 24 bits and the host part is the final 8 bits. The anonymized IP address can be anonymized with 32-bit key (11101010010011010010110110010010).

**Table 2.** An example of possible results from 5 levels

Non-Anonymization	n-Left Anonymization	n-Right Anonymization	Full Anonymization	Randomly Full Anonymization
161.200.93.0	75.133.112.0	161.200.93.146	75.133.112.146	50.204.154.136
161.200.93.1	75.133.112.1	161.200.93.147	75.133.112.147	151.33.86.11
161.200.93.2	75.133.112.2	161.200.93.144	75.133.112.144	192.37.246.138
161.200.93.3	75.133.112.3	161.200.93.145	75.133.112.145	91.154.158.81
161.200.93.4	75.133.112.4	161.200.93.150	75.133.112.150	251.28.175.177
161.200.93.5	75.133.112.5	161.200.93.151	75.133.112.151	238.131.107.78
161.200.93.6	75.133.112.6	161.200.93.148	75.133.112.148	66.126.74.200
161.200.93.7	75.133.112.7	161.200.93.149	75.133.112.149	253.214.9.219
161.200.93.8	75.133.112.8	161.200.93.154	75.133.112.154	37.226.102.49
161.200.93.9	75.133.112.9	161.200.93.155	75.133.112.155	226.2.206.69
161.200.93.10	75.133.112.10	161.200.93.152	75.133.112.152	108.208.137.21
161.200.93.11	75.133.112.11	161.200.93.153	75.133.112.153	117.143.52.30
161.200.93.12	75.133.112.12	161.200.93.158	75.133.112.158	60.113.106.222
161.200.93.13	75.133.112.13	161.200.93.159	75.133.112.159	14.42.186.251
161.200.93.14	75.133.112.14	161.200.93.156	75.133.112.156	203.190.237.11
161.200.93.15	75.133.112.15	161.200.93.157	75.133.112.157	245.15.120.136
161.200.93.16	75.133.112.16	161.200.93.130	75.133.112.130	104.170.90.1
161.200.93.17	75.133.112.17	161.200.93.131	75.133.112.131	158.54.30.228
161.200.93.18	75.133.112.18	161.200.93.128	75.133.112.128	192.22.47.232
161.200.93.19	75.133.112.19	161.200.93.129	75.133.112.129	27.116.57.62
161.200.93.20	75.133.112.20	161.200.93.134	75.133.112.134	83.188.20.57

From Table 1, the result levels 1 to 4 are consistent in prefix-preserving anonymization; they can preserve the network prefix among the same network domain from original IP addresses. The result of the last level is consistent in non-prefix-preserving anonymization. However, this level of anonymization is used to analyze the results that do not require prefix-preserving such as reporting to the public.

The benefits of this scheme are following as;

1. It appropriates for two organizations which have different views in exchanging network data by using privacy levels
2. It appropriates for packet tracers and sniffers, and suite for real-time high speed network. Because it can anonymize some bits or parts of IP address. This can reduce the computation times.

## CONCLUSION AND FUTURE WORKS

In this paper, we propose 5 levels of privacy for IP address anonymization: non-anonymization, n-left anonymization, n-right anonymization, full anonymization, and randomly full anonymization. We apply these privacy levels to prefix-preserving anonymization, specifically to Crypto-PAn. Our scheme anonymizes the necessary bits or parts of IP address by considering different privacy levels and views. It benefits any organizations in exchanging network data and also appropriates for packet tracers and sniffers.

Our future works are defining the factors concerning IP address structure, network analysis functions, and cyber laws to consider and select the appropriate privacy levels for our anonymization scheme and combining such factors by using rule-based method.

## REFERENCES

1. G. Minshall, *TCPdpriv Command Manual*, July 1996.
2. J. Xu, J. Fan, M. H. Ammar, and S. B. Moon, *Prefix-preserving IP Address Anonymization: Measurement based Security Evaluation and a New Cryptography based Scheme*, IEEE International Conference on Network Protocols (ICNP), 2002, 280-289.
3. Q. Zhang and X. Li, *An IP Address Anonymization Scheme with Multiple Access Levels*, Lecture Notes in Computer Science (LNCS), Springer-Verlag Berlin/Heidelberg, International Conference on Information Networking (ICOIN), 2006, 793-802.
4. R. Ramaswamy, T. Wolf, High-Speed Prefix-Preserving IP Address Anonymization for Passive Measurement Systems, *IEEE/ACM Transactions on Networking (TON)*, 2007, **15**(1), 26-39.
5. J.F. Kurose and K.W. Ross, *Computer Networking: A Top-Down Approach Featuring the Internet*, 2nd Edition, Addison- Wesley Publishing Company, New York, 2003, 331-342.
6. M. Peuhkuri, *A Method to Compress and Anonymize Packet Traces*, ACM SIGCOMM Internet Measurement Workshop, 2001, 257-261.
7. Q. Zhang, J. Wang, and X. Li, *On the Design of Fast Prefix-Preserving IP Address Anonymization Scheme*, Lecture Notes in Computer Science (LNCS), Springer-Verlag Berlin/Heidelberg, 6th International Conferences on Information, Communications and Signal Processing (ICICS), 2007, 177-188.

## ACKNOWLEDGMENTS

---

The financial support from Thailand Graduate Institute of Science and Technology (TGIST) is gratefully acknowledged. The scholar ID is TG-44-09-50-076M and the grant number is TGIST 01-50-076.

---