Payload analysis of anonymous communication system with host-based rerouting mechanism

Hongfei Sui Jianer Chen Songqiao Chen Jianxin Wang
College of Information Science and Engineering
Central South University
ChangSha, 410083, P. R. China
Email: hfsui@sina.com, {jianer, csq, jxwang}@mail.csu.edu.cn

Abstract

Host-based rerouting mechanism is a routing scheme that stores and forwards data in application layer. With this, users can communicate in a indirect way. Thus, identity information such as IP addresses can be effectively hidden against eavesdropper. In anonymous communication systems, such as Mixes, Onion Routing, and Crowds, this mechanism is adopted to provide anonymity. This mechanism, however, can result in extra overhead in performance such as communication delay and participant payload, which may affect the applications of anonymous communication systems. In this paper, we study quantitatively the participant payload induced by host-based rerouting mechanisms. A probability formula for calculating the participant payload is derived, which shows that the participant payload is determined by the number of participants, the number of rerouting paths, and the probability distribution of the length of rerouting paths. Applying this formula to the practical anonymous communication system, Crowds, we get immediately the precise expected participant payload, which significantly improves Reiter and Rubin's original analysis and demonstrates that the participant payload in Crowds remains a constant and independent of the variation of the number of participants in Crowds. Simulation results are presented to testify our theoretical analysis.

1. Introduction

Anonymous communication provides protection for identity information of communication participants, e.g. IP address. Anonymity could be categorized into three types [12]: *Sender anonymity* means that a particular message is not linkable to any sender and that to a particular sender, no message is linkable. *Recipient anonymity* means that a particular message cannot be linked to any recipient and that to a particular recipient, no message is linkable. *Relationship anonymity* means that it is untraceable who communicates with whom, i.e. sender and recipient (or recipients in case of multicast) are unlinkable. Recent researches mainly focus on sender anonymity.

Current anonymous communication systems include DC-Net [5, 19], Mixes [4, 11], Anonymizer [1], Anonymous Remailer [2], LPWA [7], Onion Routing I [9, 16, 18, 17], Onion Routing II [17], Crowds [13, 14], Hordes [3], Freedom [8], and PipeNet [6]. All these systems adopt host-based rerouting mechanism and/or traffic padding to provide anonymity. Host-based rerouting mechanism is a routing scheme applied in application layer. It provides indirect communication for users. Hosts involved in a communication store and forward data in application layer, thus form a virtual path consists of multiple security channels, which is called rerouting path. By simply checking the source and destination address in the header of IP packets transferred in channel, eavesdropper outside could not get the real IP address of sender and/or recipient. Even the communication recipient could not get the real IP address of the sender. Therefore, identity information of communication participants is hidden. Anonymous communication systems with host-based rerouting mechanism usually provide sender anonymity and relationship anonymity. For example, Mixes provides sender anonymity in e-mail. Onion Routing provides relationship anonymity for real-time communication. And Crowds provides sender anonymity for Web browsing to hide identity information in HTTP request which is usually exploited by Web site to get browser's identity.

The host-based rerouting mechanism brings as well extra overhead, such as communication delay, and payload on system participant. This need to be analyzed quantitatively in theory in order to make a tradeoff. Guan et al. [10] measured anonymity by information entropy, and investigated how the capability of anonymous communication system

Proceedings of the Eighth IEEE International Symposium on Computers and Communication (ISCC'03)
1530-1346/03 $17.00 © 2003 IEEE

be affected by the length of rerouting path. Reiter and Rubin [13] calculated payload on each participant of Crowds system approximately. Wright et al. [21] presented a comparative analysis about the anonymity and overhead of several anonymous communication system. And Wang et al. [20] improved the rerouting algorithm to limit the length of rerouting path effectively. This article investigates participant payload in anonymous communication system with host-based rerouting mechanism. We start by a study of the model of host-based rerouting mechanism. Then, we derive a probability formula for calculating payload on each participant in anonymous communication systems with host-based rerouting mechanism. Combining the probability formula with the length control strategy of Crowds anonymous Web browsing system [13], we derive immediately the precise participant payload of Crowds. To testify the analysis result, we also run a simulation.

2. Anonymous communication system with host-based rerouting mechanism

Guan et al. [10] presented a model of anonymous communication system with hosted-based rerouting mechanism. In this section, we describe this model and introduce some new conceptions which is necessary for analysis.

2.1. System model

Anonymous communication system with host-based rerouting mechanism can be viewed as a multi-proxy communication system which stores and forwards data to provide anonymity protection. To simple the analysis, we focus on the sender anonymity protection. The case of relationship anonymity is similar to it. A anonymous communication system with host-based rerouting mechanism consists of a set of hosts, say $V = \{v_j | 0 \leq j < N\}$, of which the element v_j is called participant, the number of participants is $|V| = N(N \geq 1)$. N is fixed during a run interval, e.g. an hour. By security channels, participants can communicate with each other directly. User requiring anonymous communication service chooses a participant $s \in V$ as the proxy, and passes the address of recipient to it. The proxy initiates a rerouting path consists of multiple participants towards the recipient, thus establishes a indirect communication between user and recipient. Formally, a rerouting path

$$\tau =< s, I_1, I_2, \cdots, I_t, \cdots, I_L, r >$$

consists of a sender $s \in V$, a recipient $r \notin V$, and intermediators $I_t (I_t \in V, 1 \leq t \leq L)$ which are participants in the system. Here we view the participant chosen for the proxy of user as the sender. $L(L = 1, 2, \cdots)$ is the number of intermediators on the rerouting path which is called length of rerouting path and conforms to the probability distribution

$$Pr\{L = k\} = f(k)$$
$$(0 \leq f(k) \leq 1, \sum_{k=1}^{\infty} f(k) = 1, k = 1, 2, \cdots)$$

A rerouting path is maintained for the communication in a running interval (e.g. an hour). During a running interval, multiple rerouting paths can be established in a system. Let $P(P = 1, 2, \cdots)$ be the number of rerouting paths in system. A forwarding task denotes the event that a participant acts as a intermediator on a rerouting path. Let payload F_j be the number of forwarding tasks on a participant v_j, which is equal to the number of appearance the participant makes as intermediator on all rerouting paths. Figure 1 shows an anonymous communication system with host-based rerouting mechanism. As depicted in the figure, the number of participants $N = 16$, and the number of rerouting paths $P = 2$. Two rerouting paths, $\tau_1 =< 0, 5, 2, 7, 11, 8, r_1 >$ and $\tau_2 =< 5, 10, 3, 9, r_2 >$, are established. Participant 0 and 5 are the sender of τ_1 and τ_2 respectively. The length of rerouting paths are $L_1 = 5$ and $L_2 = 3$ respectively. Participant 10 takes on one forwarding task, i.e. $F_{10} = 1$. Also there is only one forwarding task on participant 5, i.e. $F_5 = 1$, though it appears twice on the two rerouting paths. Note that the sender is regarded as an intermediator of rerouting path in Reiter's definition [13], which results in a slightly difference between the analysis result of this paper and that of Reiter.

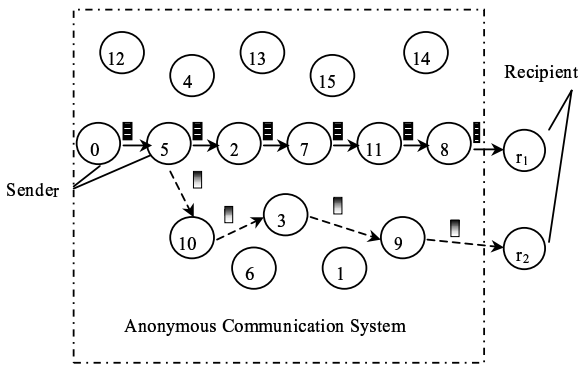

Figure 1. Anonymous Communication System with Host-based Rerouting Mechanism

2.2. Rerouting Algorithm

As described above, on constructing a rerouting path, several issues should be considered. As shown in figure 2, the rerouting algorithm includes two steps typically. Firstly the length of the rerouting path L should be determined. We say the strategy applied here is *length control strategy*.

Proc ReroutingPathGenerating;

Input

V, the set of hosts in anonymous communication system;

s, the sender of rerouting path;

r, the recipient of rerouting path;

Output

$\langle s, I_1, I_2, \cdots, I_L, r\rangle$, a rerouting path

Begin

Step 1: Determine the length of rerouting path L;

Step 2: Choose hosts sequence $I_1, I_2, \cdots, I_L$;

End

Figure 2. Rerouting Algorithm

Moreover, intermediators on the rerouting path should be chosen. We say the strategy applied here *member selection strategy*. It should be noted that not both the steps appear explicitly in practical system, e.g. the step for length control is omitted in Onion Routing I since the length of rerouting path is fixed to a constant.

Alternatively, there are two length control strategies: Fixed Length Strategy and Variable Length Strategy. With the Fixed Length Strategy, length of rerouting path is a constant C, i.e.

$$Pr\{L = k\} = f(k) = \begin{cases} 1, & k = C \\ 0, & k \neq C \end{cases} \quad (k = 1, 2, \cdots)$$

In Onion Routing and Freedom, Fixed Length Strategy is adopted. Opposite to Fixed length Strategy, length of rerouting path L is a discrete random variable in the Variable Length Strategy. In Crowds and Onion Routing II, the Variable Length Strategy is adopted. We discuss the case of Variable Length Strategy and regard the Fixed Length Strategy as a special case of Variable Length Strategy. Also there are two strategies can be adopted in member selection: Randomized Strategy and Non-Randomized Strategy. With the Randomized Strategy, host is chosen uniform randomly from all the N hosts (including sender itself) to be an intermediator on rerouting path. With the Non-Randomized Strategy, host is chosen out from all the N hosts according to payload, reliability of host. We discuss Randomized Strategy since it is now adopted in Crowds and Onion Routing II.

3. Analysis of Participant Payload

In this section, we calculate payload on each participant in anonymous communication system with host-based rerouting. As described above, payload F can be obtained accordingly by calculating the number of appearances that a host makes on all rerouting paths. Assume there are N hosts and P rerouting paths in running interval. Let L_m, the length of the m^{th} rerouting path τ_m $(m \leq P)$, is an discrete random variable conforming to probability distribution

$$P\{L_m = k\} = f(k)$$
$$(0 \leq f(k) \leq 1, \textstyle\sum_{k=1}^{\infty} f(k) = 1, k = 1, 2, \cdots) \quad (1)$$

Consider a arbitrary participant $v_j \in V$. Let F_j be the participant payload of v_j. Let R_j be the number of appearances that participant v_j makes on all the rerouting paths. Let R_j^m be the number of appearances that participant v_j makes on τ_m. We get

$$F_j = R_j = \sum_{m=1}^{P} R_j^m \quad (2)$$

According to equation (1), we get $E(L_m)$, the expected length of the rerouting path τ_m

$$E(L_m) = \sum_{k=1}^{\infty} kPr\{L_m = k\} = \sum_{k=1}^{\infty} kf(k) \quad (3)$$

Suppose that the Randomized Strategy is applied for member selection, participant for intermediator on rerouting path is chosen out uniform randomly from all the N participants in system. We can get the conditional probability that participant v_j appears i times on rerouting path of which the length is k

$$Pr\{R_j^m = i \mid L_m = k\} = C_k^i (\frac{1}{N})^i (\frac{N-1}{N})^{k-i}$$
$$i = 0, 1, 2, \cdots, k \quad (4)$$

According to equation (1)(4), the probability that length of τ_m is k and participant v_j appears i times on τ_m is

$$\begin{aligned} &Pr\{R_j^m = i, L_m = k\} \\ &= Pr\{L_m = k\} Pr\{R_j^m = i | L_m = k\} \\ &= f(k) C_k^i (\frac{1}{N})^i (\frac{N-1}{N})^{k-i} \end{aligned} \quad (5)$$

So we can get the probability that participant v_j appears i times on τ_m

$$\begin{aligned} &Pr\{R_j^m = i\} \\ &= \sum_{k=1}^{\infty} Pr\{R_j^m = i, L_m = k\} \\ &= \begin{cases} \sum_{k=1}^{\infty} f(k) (\frac{N-1}{N})^k, & i = 0 \\ \sum_{k=i}^{\infty} f(k) C_k^i (\frac{1}{N})^i (\frac{N-1}{N})^{k-i}, & i \geq 1 \end{cases} \end{aligned} \quad (6)$$

According to equation (2)(3)(6), we can get the following theorem.

Theorem 1: In anonymous communication system with host-based rerouting mechanism, there are $N(N = 1, 2,)$ participants and $P(P = 1, 2,)$ rerouting paths in an running interval. The length of the $m^{th}(1 \leq m \leq P)$ rerouting path $\{L_m\}$ is independent discrete random variable, and conforms to probability distribution

$$P\{L_m = k\} = f(k)$$
$$(0 \leq f(k) \leq 1, \sum_{k=1}^{\infty} f(k) = 1, k = 1, 2, \cdots)$$

Randomized Strategy is adopted for constructing rerouting path. Then the expectation of F_j, payload on a arbitrary participant $v_j(0 \leq j < N)$

$$E(F_j) = (\frac{P}{N})E(L_m) \tag{7}$$

where $E(L_m)$ is the expected length of rerouting path.

See appendix for the proof of Theorem 1. Theorem 1 demonstrates that in anonymous communication system with host-based rerouting mechanism, participant payload is determined by the number of participants in the system N, the number of rerouting paths P, and the expected length of rerouting path $E(L_m)$. With a specific length control strategy, the expected participant payload $E(L_m)$ stays to be a constant, since the probability distribution that the length of rerouting path conforms to is determined. In this case, the participant payload is mainly determined by N and P. Participant payload could become over-heavy for participant, if the increase on the number of participants N is limited while P could increase unlimitedly. This is usually a case in practice. Therefore, number of rerouting paths P should be limited to keep participant payload low in practice. We will see in following section that means is adopted to limit P to a maximum value N in Crowds, which makes the maximum participant payload to be a constant in the worst case.

4. Participant Payload in Crowds

In this section, we apply Theorem 1 to the analysis of a practical system - Crowds. Since the length control strategy in Onion Routing II is the same as that in Crowds, participant payload in Onion Routing II is similar to that in Crowds.

Crowds is a anonymous communication system with host-based rerouting mechanism [13, 14]. It provides sender anonymity for web browsing. Host need to be anonymous must join in Crowds as participant, to provide protection for other participant as well as to be protected. When a participant needs to initiate an anonymous communication, it sends its request to another participant. On receiving the request, the other participant forwards the request to the next participant or submits the request to the responder on behalf of the communication initiator. Specifically, a proxy named Jondo is run on the host, which forwards all HTTP requests from the local browser and that from Jondo running on other participants. Initially, the Jondo registers itself in the Blender, a host which manages all Jondos in Crowds, obtains Jondo list and key list from the Blender. On receiving the first HTTP request from the local browser, the Jondo chooses a Jondo uniform randomly from the Jondo list as its successor and forwards the request to the successor. The successor forwards the request to another Jondo with probability $P_f(1/2 \leq P_f < 1)$, or submits the request to end server. Thus, a rerouting path consists by Jondos is formed. Subsequent requests from the local browser will be forwarded along the rerouting path. Figure 3 shows a Crowd system including 6 participants and 6 rerouting paths, i.e. $N = 6$ and $P = 6$. The rerouting paths are: $\langle 1, 5, server1\rangle$; $\langle 2, 0, 2, server2\rangle$; $\langle 3, 1, 0, server3\rangle$; $\langle 4, 4, server4\rangle$; $\langle 5, 4, 0, server1\rangle$; $\langle 0, 3, server2\rangle$.

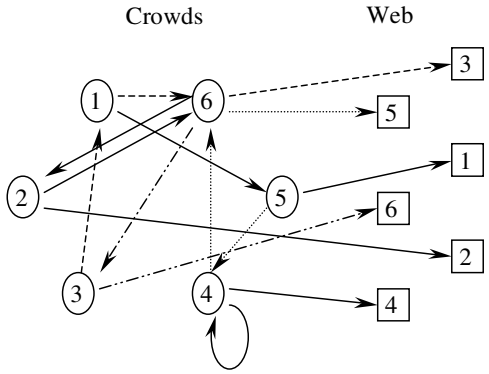

Figure 3. Crowds

As described above, Variable Length Strategy and Randomized Strategy are adopted in Crowds. During a running interval, there are P rerouting paths be initiated ($P \leq N$). We focus on the case that there are maximum number of rerouting paths in the system, i.e. $P = N$. According to the length control strategy in Crowds, the length of rerouting path conforms to probability distribution

$$P\{L_m = k\} = f(k) = (1 - P_f)P_f^{k-1} \tag{8}$$
$$(0 \leq f(k) \leq 1, \sum_{k=1}^{\infty} f(k) = 1, k = 1, 2, \cdots)$$

we can get the expected length of rerouting path

$$E(L_m) = \frac{1}{1 - P_f}, (\frac{1}{2} \leq P_f < 1) \tag{9}$$

Since $P = N$, we can get $E(F_j)$, i.e. expected participant payload of v_j according to Theorem 1

$$E(F_j) = (\frac{P}{N})E(L_m) = (\frac{1}{1-P_f}) \qquad (10)$$

It should be noted that the sender is regarded as an intermediator on rerouting path in Reiter's analysis. This results the expected participant payload on v_j to be $\frac{1}{1-P_f}+1$, since v_j acts as sender on all rerouting paths just one time. According to theorem 7.1 in Reiter[13], expected participant payload is bounded from above by $\frac{2n}{(n-1)(1-P_f)^2}$. Obviously, our analysis improves Reiter's analysis result and presents payload on host precisely.

Furthermore, we tested the participant payload by simulation and calculated average payload F_a to verify the analysis result. In each interval during the runtime, N rerouting paths is established in the simulation system. For every N or P_f the system is run about 100,000 interval. Participant payload in each interval is logged and the average participant payload F_a is calculated. Figure 4 and 5 show variation of average participant payload F_a when N and P increase respectively. In contrast, result calculated from expected payload by this paper and the upper bound of expected payload on host by Reiter are also shown in figures. In figure 4, variation of average payload on host F_a is a horizontal line which is coincide with $E(F) = 6$, the result calculated from the participant payload by this paper $\frac{1}{1-P_f}+1$, and much lower than Reiter's upper bound. In figure 5, average participant payload F_a increases slowly when P_f increases, which is also coincide with the analysis result by this paper. Therefore, our analysis result is correct.

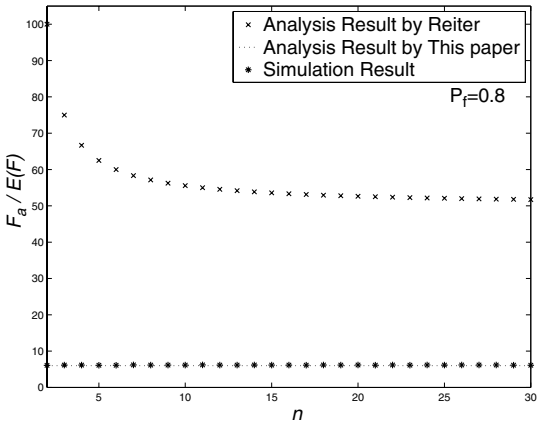

Figure 4. Payload VS Number of participant

In addition, we have following conclusion. First, variation of the number of participants N and the number of rerouting path P does not affect the participant payload F. The expected participant payload $E(F)$ is simplified to be a function of P_f since the number of rerouting paths P is limited to be less than the number of participants N in Crowds. During the running interval, P_f is kept to be constant which

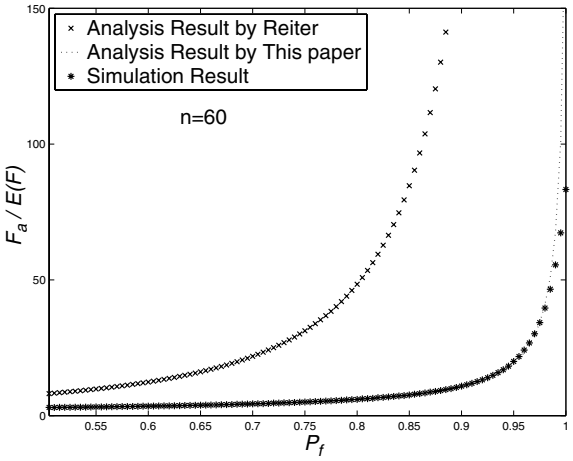

Figure 5. Payload VS Probability of forwarding

results in the expected payload on participant $E(F)$ to be a constant correspondingly. That means, dynamic variation of the number of participants N does not result in variation of the participant payload essentially. Thus, Crowds scales well. Second, expected participant payload in Crowds is determined uniquely by P_f. Increasing P_f may lead to longer rerouting path and heavier participant payload. Therefore, P_f should be chosen carefully in practice to obtain better performance. We refer the interested readers to [15] for more details.

5. Conclusion

In this paper, we investigate underlying host-based rerouting mechanism of anonymous communication system, and derives a probability formula for calculating expected participant payload. The probability formula demonstrates that participant payload is determined by number of rerouting paths, probability distribution of the length of rerouting path, and number of hosts in anonymous communication system with host-based mechanism. The number of rerouting paths in a system should be limited in case the participant payload becomes too high. Then we analyze participant payload in Crowds with the probability formula, and calculate that expected participant payload in Crowds is $\frac{1}{1-P_f}+1$. This result is verified by simulation and improves Reiter's analysis. It demonstrates that participant payload in Crowds is kept to be a constant when system is running with maximum payload while independent of variation of the number of hosts in system. Thus, Crowds scales well.

6. Acknowledgment

This research is supported in part by the Major Research Plan of National Natural Science Foundation of China, Grant No. 90104028.

References

[1] The anonymizer. *http://www. anonymizer. com*.

[2] Anonymous remailer. *http:// www.lcs.mit.edu /research /anonymous.html*.

[3] B. N. L. C. Shields. Hordes: a protocol for anonymous communication over the internet. *ACM Journal of Computer Security*, 10-3, 2002.

[4] D. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 24-2:84–88, 1981.

[5] D. Chaum. The dinning cryptographers problem: Unconditional sender and recipient untraceabilit. *Journal of Cryptology*, 1/1:65–75, 1988.

[6] W. Dai. Pipenet 1.1. *Http://www.eskimo.com /weidai /pipenet.txt*.

[7] E. Gabber, P. Gibbons, Y. Matias, et al. How to make personalized web browsing simple, secure, and anonymous. *Lecture Notes in Computer Science*, 1318:17–31, 1997.

[8] Goldberg and A. Shostack. Freedom network 1.0 architecture and protocols. *http:// www.freedom.net /info /freedom-papers /index.html*, 1999.

[9] D. Goldschlag, M. Reed, and P. Syverson. Onion routing for anonymous and private internet connections. *Communications of the ACM*, 42-2:39–41, 1999.

[10] Y. Guan, X. Fu, R. Bettati, et al. An optimal strategy for anonymous communication protocols. *Proceedings of the 22th International Conference on Distributed Computing Systems*, July 2002.

[11] C. Gulcu and G. Tsudik. Mixing email with babel. *Proceedings of the 1996 Symposium on Network and Distributed System Security*, 1996.

[12] A. Pfitzmann and M. K?hntopp. Anonymity, unobservability, and pseudonymity - a proposal for terminology. *Proc. Workshop on Design Issues in Anonymity and Unobservability. LNCS 2009*, 99:1–9, 2001.

[13] M. K. Reiter and A. D. Rubin. Crowds: Anonymity for web transactions. *ACM Transactions on Information and System Security*, 1-1:66–92, November 1998.

[14] M. K. Reiter and A. D. Rubin. Anonymous web transactions with crowds. *Communications of the ACM*, 42(2):32–38, 1999.

[15] H. Sui, J. Wang, J. Chen, et al. An analysis of forwarding mechanism in crowds. *Proc. IEEE International Conference on Communications (ICC'2003), to appear*, 2003.

[16] P. Syverson, D. Goldschlag, and M. Reed. Anonymous connections and onion routing. *Proceedings of the IEEE Symposium on Security and Privacy, IEEE CS Press*, pages 44–54, May 1997.

[17] P. Syverson, M. Reed, and D. Goldschlag. Onion routing access configuration. *DISCEX 2000: Proceedings of the DARPA Information Survivability Conference and Exposition, Hilton Head, SC, IEEE CS Press*, pages 34–40, 2000.

[18] P. Syverson, G. Tsudik, M. Reed, et al. Towards an analysis of onion routing security. *Workshop on Design Issues in Anonymity and Unobservability*, July 2000.

[19] M. Waidner. Unconditional sender and recipient untraceability in spite of active attacks. *Eurocrypt89*, (7), April 1989.

[20] W. Wang, J. Chen, J. Wang, et al. A anonymous communication protocol based on groups with definite route length. *Journal of computer research and development(In Chinese)*, 24(5):463–467, 2003.

[21] M. Wright, M. Adler, B. N. Levine, et al. An analysis of the degradation of anonymous protocols. *Proc. ISOC Network and Distributed system Security Symposium (NDSS)*, 2002.

7. Appendix

Proof of Theorem 1

Proof: According to equation(6), we can get expected R_j^m

$$\begin{aligned} R(R_j^m) &= \sum_{i=0}^{\infty} iPr\{R_j^m = i\} = \sum_{i=1}^{\infty} iPr\{R_j^m = i\} \\ &= \sum_{i=1}^{\infty} i \sum_{k=i}^{\infty} f(k) C_k^i (\frac{1}{N})^i (\frac{N-1}{N})^{k-i} \\ &= \sum_{k=1}^{\infty} f(k) \sum_{i=1} k i C_k^i (\frac{1}{N})^i (\frac{N-1}{N})^{k-i} \quad (11) \end{aligned}$$

By Binomial Theorem, we can get

$$\sum_{i=0}^{k} C_k^i a^{k-i} (bx)^i = (a+bx)^k \quad (12)$$

We can get

$$\sum_{i=0}^{k} i C_k^i a^{k-i} b^i x^{i-1} = kb(a+bx)^{k-1} \quad (13)$$

Let $x = 1$,$a = \frac{N-1}{N}$,$b = \frac{1}{N}$, then equation(13) is

$$\sum_{i=1}^{k} i C_k^i a^{k-i} b^i x^{i-1} = kb(a+bx)^{k-1} \quad (14)$$

According to equation (3)(11)(14), we can get

$$\begin{aligned} E(R_j^m) = \sum_{k=1}^{\infty} f(k)(\frac{k}{N}) = \quad & (\tfrac{1}{N}) \textstyle\sum_{k=1}^{\infty} f(k) k \\ = \quad & (\tfrac{1}{N}) E(L_m) \quad (15) \end{aligned}$$

According to (2)(15), we can get

$$\begin{aligned} E(F_j) = E(R_j) = E(\sum_{m=1}^{P} R_j^m) = \quad & \textstyle\sum_{m=1}^{P} E(R_j^m) \\ = \quad & (\tfrac{P}{N}) E(L_m) \quad (16) \end{aligned}$$

End proof