Research on Anonymous Network Topology Analysis

Shiyu Kang

Guangxi Vocational & Technical Institute of Industry, Nanning Guangxi, China 53001

Keywords: anonymous communication, node selection, DHT, network topology

Abstract. In response to protect the privacy of Internet users, a variety of privacy enhancing technologies (PETs) have emerged. As one of privacy enhancing technologies, anonymous communication has been extensively studied from various aspects by researchers. In this paper, we investigate existing organizations and universities which study anonymous communication from the perspective of network topology, and related projects. Then we survey related papers to anonymous communication published in recent years, which focus on the analysis of node selection (especially in the Tor). Finally, some related problems and follow-up study are presented to be studied deeply in future.

Introduction

Nowadays, Internet privacy becomes more and more important and sensitive, and has been one of the latest buzz words to hit the Internet world. In response to protect the privacy of Internet users, a variety of privacy enhancing technologies (PETs) has emerged, such as anonymous communication. Anonymity set includes senders' anonymity set and receivers' anonymity set. The former set can disjoint, overlap, and intersect the latter set. If the size of anonymity set is bigger, and the probability of sending is more similar to the probability of receiving, then the anonymity is stronger.

At present, anonymous communication systems which have been developed varies widely from node discovery, node selection to transporting messages. The typical systems include Tor, Crowds [1], Tarzan, I2P, and so on. In particular, Tor is the most popular and widespread system, which is the object of numerous academic research and testing platform in anonymous communication. Existing work focuses on the research of these systems from various aspects and challenges, especially addressing the key issues of anonymity and performance in communication. Moreover, many anonymous systems (such as Onion Routing, Crowds, Freedom, MorphMix and Tarzan) have achieved based on P2P, while these systems are more dynamic in nature for many nodes (called relays) are in the network for a short time. When a new node enters the network, it needs other nodes in the network to build path. And when the node exits network, the user in the path has to build a new path. One problem is the node joining and exiting also need the other nodes in the network, while the changing anonymous set make the problem more difficult. Meanwhile, node performance is different from each other. Based on barrel principle, it may cause the problem that poor performance of node can degrade the efficiency of anonymous path, even if the performance of other nodes on the path is very well. To summarize, how to select node is crucial to build anonymous path.

Based on the above analysis, this paper firstly introduces and analyzes major organizations and universities which study anonymous communication from topology and node discovery. We present a framework of anonymous network topology, and a visual illustration of this analysis that shows the progression of the researches of network topology and node discovery in anonymous communication.

Recruit new relays

During the process of anonymous communication, there are many problems, such as dynamic joining and leaving of nodes. Frequent node churn make it difficult to maintain a stable network topology. Moreover, anonymity set consists of numerous nodes, and the size of anonymity set has a

direct effect on the anonymity and performance of system. For example, Tor is made of voluntary node, its performance depend on its volunteers to donate their resources. However, the resource can't meet the ever-increasing demand on Tor. Joining new node can solve the problem, because new node expands the diversity and size of the anonymity set to improve system performance and anonymity. But the sufficient incentives for volunteers to run relays are currently lacking.

In recent years, various incentives to join anonymous network is proposed. Back in 2008, Androulaki et al. [2] introduced a hybrid payment system that combined the Peppercoin Micropayment system with "one use" electronic cash, proposed one theoretical design mechanism, Payment for Anonymous Routing (PAR), which can provide economic incentives for network participants. During the mechanism, a centralized bank released coins to clients, clients should pay coins when using every circuit.

In 2010, Ngan et al. [3] proposed "gold star" mechanism, that the Tor directory servers measure the performance of relays and grant the satisfactory relays in the directory with "gold star", that the better performance the relays have, the higher chance the relays are selected to build circuit.

Later, Jansen et al. [4] proposed BRAIDS, which encourage users to run Tor relays by introducing relay-specific tickets for service accounting. The tickets are embedded into Tor cells to request some sort of service，such as low-latency service, high-throughput service, and normal service.

Further, unlike above incentive and e-cash approaches that need centralized banks, Moore et al. [5] proposed decentralized incentive scheme, which Tortoise enforced low universal rate limit on individual client connections at the network's ingress points to achieve speedup. Recently, unlike recruiting relays, Wang et al. [6] propose user reputation system, rBridge, which uses an introduction-based mechanism to recruit new relays.

Build anonymous path

Although one node join the anonymous system，it does not mean the node must be used, but open the possibility of building anonymous path. The most section in building the path is anonymous routing algorithm. Anonymous routing algorithm is the core of anonymous communication systems, because it determines the whole system performance and security. Furthermore, the number of users decides anonymity degree that anonymous system can offer, while whether users use the anonymous system depend on system performance, so it's necessary to propose better routing algorithms to design better anonymous systems.

Before building the path, it's necessary to determine the anonymous path length, which is subject to network topology, and can be divided into variable length path and fixed length path. For variable length path (such as Crowds [1]), sender only selects the first node, and can't determine the path length. For example, Crowds use the forwarding probability way that each node select next node to forward message or directly send the message to receiver depending on the forwarding probability. Thus the path length varies from each node forwarding probability. In extreme circumstances, the length may be infinitely long, then the latency will be incalculable.

Node selection

Routing algorithm can divide into two sections: how many nodes and which nodes can be used to build anonymous path. In other words, they can be called node selection and path construction. Node selection can affect the bandwidth utilization ratio and throughput, and mainly solves the problem that how safely, rationally and efficiently select relay node. Path construction mainly solves the problem that how to use the selected nodes quickly and effectively construct path. During node selection algorithms, different proposals have been presented from varying elements. Broadly speaking, node selection can divide into two classes: based on node characteristics (such as bandwidth) and based on link characteristics (such as latency). But the classification can't summarize all existing selection algorithms. Considering many network metrics can describe the

performance of an anonymous path, such as bandwidth, latency, network jitter, loss, etc. existing algorithms can be subdivided into five classes: bandwidth based, latency based, as-aware, application-aware and trust based methods.

Bandwidth-based node selection

Many researchers have attempted to improve the performance problem. For example, Snader and Borisov propose an opportunistic bandwidth measurement algorithm to improve Tor selection algorithm. Moreover, they also propose a tunable path selection that different users vary different sets of node from their preferences for anonymity versus performance. Then, Murdoch et al. explore Tor current path selection algorithms with one Tor path simulator, such as bandwidth-weighted algorithm and uniform selection path algorithm.

Latency-based node selection

Considering that node's actual bandwidth is unknown, in other words, it's difficult to prove the node's self-reported bandwidth is true, bandwidth–based selection can cause some problems. Moreover, some nodes in the anonymous system are in a constant state of congested or unused, so that it not only leaves loopholes for attackers, but also degrades network performance. To solve these problems, latency-based path selection has been presented. Latency in anonymous path includes transmission delay, queuing delay and propagation delay. Transmission delay refers to the amount of time required to push all of the packet's bits into the wire. Queuing delay is the delay caused by a packet that it may wait in a queue until handled. Propagation delay, that is link latency, is the amount of time it takes for the packet to travel from the sender to the receiver along the anonymous path. Moreover, latency is the selection criteria of linked-based path selection. Put another way, the lower the delay is, the higher priority the link that is selected have.

In 2009, Panchenko and Renner propose novel path selection based on actively measured the latencies of anonymous paths in terms of round-trip times (RTTs), and using passive observations of throughput to estimate available link-wise capacities.

Sherr et al. present link-based relay selection which can provide more flexible routing and anonymous path with low latency and network jitter, and also introduce virtual coordinate system. The Euclidean distance between nodes in the virtual coordinate system is regarded as metric of latency. To protect the anonymity, both parties of communication don't take part in virtual coordinate system, only the set of nodes do. Before selecting path, sender firstly calculates the nodes' distance in the coordinate system, then estimate the latency of potential path.

In 2012, Wang et al. points out the neglected problem easily that Tor path selection algorithm don't consider the current load of nodes. Unlike redesigning some congestion control methods, they lay emphasis upon identifying and avoiding congested nodes. Congestion-aware path selection algorithm has come up that avoid selecting very congested circuits to decrease node latency. In detail, to decrease latency as an indicator of congestion, sender firstly use a combination of lightweight active and opportunistic measurements that means opportunistically sample RTTs across potential path and subtract the lowest recorded RTT to obtain the overall latency of the path, infer node latency to judge the state of nodes whether appear congested, and select nodes which don't appear congested.

Until recently, Wacek et al. improve the Sherr's algorithm that combine Tor algorithm with Sherr et al. algorithm. Clients still use the Tor's relay selection to select k candidate paths, and then estimate the latency of each k candidate paths, finally select the path with lowest latency as the ultimate path. In addition, they indicate that if the value of k is 3, it can provide the best trade-off of performance and the time spent identifying the best path.

AS-aware node selection

Low-latency anonymous system is more vulnerable to timing attacks and statistical correlation attack. To mitigate the attack and improve anonymity, one method is location diversity by making each path more complex, such as spreading over multiple jurisdictions, not in the same autonomous system (AS). Therefore, the aware of autonomous system should be taken into account during node selection.

As is a network or a collection of networks under mutual administration that shares the same routing methodology, and can independently operate network. In 2004, Feamster and Dingledine consider location independence metric, and argue that node selection algorithms that look only at IP prefixes likely inefficiently implement location independence, in which selected paths are subject to compromise anonymity by a single AS. They have found that different node selection algorithm help minimize the chance that entry path and exit path traverse the same AS, and the longer mix path has, the smaller parts a single AS can observe.

In 2009, Edman and Syverson further study path inference algorithms, suggest that although Tor's node is increasing continuously, the probability of AS observing the ends of path is higher than previously thought. In turn, AS-aware node selection algorithm is proposed to resist AS level attackers.

In 2012, Akhoondi et al. present tunable node selection algorithm that user can set a value between 0 to 1 to balance the anonymity and latency. Moreover, they estimate the geographic locations of clients and nodes using an IP geolocation database to decrease the latency. Besides, the algorithm can predict Internet routing between nodes and clients to resist correlated attack.

Application-aware node selection

To meet some specific applications' requirements, clients should have application-aware to select nodes. In 2007, Sherr et al. firstly propose A3 to meet application-specified criteria. Later, A3 have been implemented by them, and they design a declarative language (A3LOG) that make applications to compactly specify path instantiation and node selection.

Trust-based node selection

In anonymous network, node frequently enter and exit the network, so that the trust between nodes become more and more important because it is often associated with the user status and information privacy protection. Trust-based path selection is taking trust into account selecting nodes.

In 2009, Johnson and Syverson regard the probability that the adversary fails to compromise nodes as trust. They propose node trust model, and first explicitly use trust to redesign node selection strategies, but only consider correlation attacks then. Later, they further identify the importance of trust, consider different users have different distributions on trust, then propose a novel trust-based node selection algorithm which can protect anonymous system from attacking a significant fraction of the network.

Besides these above node algorithms, the simplest algorithm is uniform random selection that the node is selected from a set of candidate nodes with uniform probability. In short, the probability that each node is selected to build anonymous path is same.

Locate random relays

Locate random relay also refers to secure lookup node. Existing locating relay methods are largely based on various distributed hash tables (DHT). DHT provides a lookup service similar to a hash table. In other words, DHT can extract some information in the file (usually a filename) as key to be unique values by hash function, and stores the (key, value) pairs, then any participating node can efficiently retrieve the value associated with a given key. Typical DHT algorithms include Chord, CAN, Pastry, Kademlia [7], and Tapestry. It's important to locate random nodes in anonymous communication systems. Unlike Tor, there are some anonymous communication systems which are based on Distributed Hash Table (DHT), such as Salsa, AP3, NISAN, Torsk, Bifrost, Cashmere ShadowWalker, and Octopus.

Based on Chord [8], there are NISAN, Salsa, Bifrost which have implemented. Salsa uses a specifically designed secure lookup over a custom DHT to select nodes, which based on a Chord-like DHT that maps nodes to a point in an ID-space corresponding to the hash of their IP address. The identities are based on hashes of the nodes' IP addresses which are organized in a tree structure. However, Salsa's lookup may suffer selective DoS. In 2009, Kondo et al. [9] design Bifrost system, which separate a node management layer (NML) realized by Chord from anonymous communication layer with multiplex encryptions to avoid anonymous route affecting

node. NML can not only use the Finger Table of Chord to search the next-hop node, but also assign backup node once relay node seceded. Based on Kademlia DHT and Myrmic [10], Torsk is designed by McLachlan et al in 2009. In Torsk, clients don't need to communicate with directory server, and combine its' lookup with root verification, buddies and cover traffic. Based on Pastry DHT, Zhuang et al. in 2005 design Cashmere, which use mix idea and multi-hop routing, and select a set of nodes in overlay namespace to solve the issues of node churn and improve the stability of anonymous path. The set of nodes is called a relay group, each node in the group can be a mix, and share a public/private key pair. Anonymous path can be built by various relay groups. AP3 is similar to Crowds, and perform a stochastic expected-length random walk. However, the lookup mechanisms in Salsa and AP3 lack anonymity, so that adversary can infer the path structure and compromise user anonymity.

Entry guard

In general, the first relay in anonymous path is often called entry node. In all relay nodes, only entry node knows the communication initiator, and malicious entry nodes can make selective DoS attack more powerful so that many anonymous paths will be compromised [11], therefore how to select entry node is of great significance to protect sender. In Tor, entry node is generally selected from node set with high stability (such as long time online) and bandwidth to cope with predecessor attack, locating hidden service, and statistical profiling. In general, clients select three entry guards to apply to all circuits, and reselect new entry guard at intervals of 30 days to 60 days.

In 2006, Øverlier and Syverson has researched entry nodes to protect hidden servers. The parameters in the nodes include the size of entry node set, performance, trust, and so on. They not only propose backup a longer list of entry nodes than normal entry node set to cope with the normal set unavailable, but also put forward layering entry nodes and select a small fixed number of nodes to always regard as entry nodes.

In 2012, Elahi et al. [12] design a simulation framework called Changing of the Guards (COGS) to provide quantitative data about Tor's entry guard selection algorithms with an empirical analysis. They suggest that natural churn and guard rotation are main factors affecting guard selection. The best balance between making guard nodes diverse and avoiding selecting malicious nodes as guard nodes should be further study.

Conclusions

In detail, there are still many problems to be solved in the topology research. Although there are many improved algorithms for node selection and entry guard proposed under the frame of theory, the practicality of these algorithms to be applied in live anonymous system is unknown, then it's urgent that develop new method to model Tor more realistically. In addition, it's worth exploring new node selection algorithm to defend themselves against various attacks, such as predecessor attack and routing capture attack. At last, some measures of improving anonymity tend to increase system loads, add long latency to the processing of service, and reduce efficiency. It's certainly worth considering that how to balance anonymous systems' anonymity and performance.

References

[1] Michael Reiter and Aviel Rubin. Crowds: Anonymity for web transactions. ACM Transactions on Information and System Security, 1(1):66{92, (1998).

[2] E. Androulaki, M. Raykova, S. Srivatsan, A. Stavrou, and S. M.Bellovin. PAR: Payment for anonymous routing. InPETS '08: Proceedings of the 8th International Symposium on Privacy Enhancing Technologies, (2008), pp. 219 - 236.

[3] Tsuen-Wan ``Johnny" Ngan, Roger Dingledine, and Dan S. Wallach, "Building Incentives into Tor", In the Proceedings of Financial Cryptography (FC '10), January(2010) .

[4] Rob Jansen, Nicholas Hopper, and Yongdae Kim, “Recruiting New Tor Relays with BRAIDS”, In the Proceedings of the 2010 ACM Conference on Computer and Communications Security (CCS 2010), Chicago, Illinois, USA, October (2010) .
[5] B. Moore, C. Wacek, and M. Sherr. Exploring the Potential Benefits of Expanded Rate Limiting in Tor: Slow and Steady Wins the Race With Tortoise. In Annual Computer Security Applications Conference (ACSAC), December (2011).
[6] Qiyan Wang, Zi Lin, Nikita Borisov, and Nicholas J. Hopper, “rBridge: User Reputation based Tor Bridge Distribution with Privacy Preservation”, In the Proceedings of the Network and Distributed System Security Symposium - NDSS'13, February (2013) .
[7] P. Maymounkov and D. Mazieres. Kademlia: A peer-to-peer information system based on the xor metric. IPTPS,(2001).
[8] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan. “Chord: a Scalable Peer-to-peer Lookup Service for Internet Applications”. In Proc. of ACM SIGCOMM, August (2001) .
[9] M. Kondo, S. Saito, K. Ishiguro, H. Tanaka, and H. Matsuo. Bifrost: A Novel Anonymous Communication System with DHT. In Second International Workshop on Reliability, Availability, and Security, (2009) , pages 324–329.
[10] Sherr, M., Loo, B.T., Blaze, M., “Towards Application-Aware Anonymous Routing”. In: USENIX Workshop on Hot Topics in Security (HotSec) August (2007) .
[11] Snader, R., Borisov, N., “A Tune-up for Tor: Improving Security and Performance in the Tor Network”. In: 15th Annual Network and Distributed System Security Symposium (NDSS) (February (2008)).
[12] Tariq Elahi, Kevin Bauer, Mashael AlSabah, Roger Dingledine, and Ian Goldberg, “Changing of the Guards: A Framework for Understanding and Improving Entry Guard Selection in Tor”, In the Proceedings of the Workshop on Privacy in the Electronic Society (WPES 2012), Raleigh, NC, USA, October (2012) .