Anonymizing Public Transportation Rides Information

Final project in course:
Advanced topics in online privacy and cybersecurity (67515)
Submitted by: Yiftach Sabag

1 Introduction

Today's public transportations systems rely significantly on computers, networking and other technologies. These systems hold and exchange large amounts of users' private information. For instance, full name, address, phone number, location, public transportation rides data, etc. This makes those transportations systems a great target for malicious attackers. We are creating a system in order to hide this data and metadata.

In Israel there are two major paying systems for public transportation:

1. "Rav-kav" cards: involve RFID technology, and close distance networking.
2. Smartphone applications: involve long-distance networking technologies and protocols.

We are targeting the second system: smartphone applications. We wish to improve these applications' privacy with a system that anonymizes the data and metadata that is sent to the Ministry of Transportation (MoT) servers.

2 Motivation For the Problem Importance

Considering only the metadata of a message: source, destination, size, time sent, with respect to public transportation, this information is extremely sensitive. Attackers that are exposed to this data can gain a great amount of information. Assuming no access to the MoT servers, Malicious individuals can learn for instance, when users use public transportation, and the approximate location of the different rides. Assuming access to the MoT servers, all the information that is linked to a user, can tell an attacker a user's public rides schedule, locations history, etc.

Different examples include:

1. Governments with access to the data may hurt its citizens' privacy, increase their control over them, and create a dark regime in the country.
2. Health-Insurance companies can profile customers, by knowing if a person's rides related to stations are near a hospital or other health service locations. By this information they can reject or approve an insurance policy.
3. Companies use targeted advertising for users, by learning their points of interest. For instance, a person that uses public transportation to go to the gym can be targeted with gym-related equipment advertisements.

4. Companies and institutions can trace whistleblowers.
5. Etc.

3 Solution Information

3.1 Key Insights

1. **What are we hiding?**
   1.1. correlation between transportation data and the identity of the user.
   1.2. correlation between messages entering the Mixnet and the messages exiting the Mixnet (going out from exit relay).
2. **From whom are we hiding it?**
   2.1. MoT
   2.2. People with access to the servers (technicians, maintenance people, etc.)
   2.3. Malicious attackers.
3. **Who do we trust?**
   3.1. We assume at least one honest relay in the Mixnet.

Note. We are OK with attackers knowing when a user is sending a message to the MixNetwork: it is a free and open-source infrastructure that anyone can use, regardless of public transportation usage. Thus, an attacker that watches alice's outgoing link, cannot know if a message sent into the Mixnet is related to a public transportation ride or not.

3.2 Attacker Model

1. We assume attacker that can monitor the user's outgoing link (messages entering the Mixnet)
2. We assume attacker that can monitor the last relay's outgoing link (messages exiting the Mixnet)
3. The attacker is passive
4. An attacker may have control over a subset of the relays inside the Mixnet, but at least one relay is honest.

3.3 Assumptions

1. Clients know public keys of relays in Mixnet.
2. Clients know the MoT server's public key.

3. At least one relay in the Mixnet is honest.
4. There are at least 3 different relays in the Mixnet chain.
5. The core message is encrypted with the MoT server's public key. (asymmetric encryption).
6. Personal data of users is either saved locally on the user's side, and is never sent to any server, or stored encrypted on the application server, without the ability to decrypt it by someone other than the user (similar to different cloud services). Credit card information can be stored encrypted on the server side, according to SSL standards.

3.4 Overview of the system

Techniques. The system uses the following techniques:

1. Mixnet
2. Asymmetric encryption (Diffie Hellmen key exchange)
3. Symmetric encryption

Encrypted data. The encrypted data that is sent to the MoT servers for statistics usage includes:

1. Line number
2. Public transportation operator
3. Travel code (related to the price of the ride)
4. Approximate boarding time (5 minutes segments)
5. Boarding station
6. destination station

Onion routing. Messages are encrypted using onion routing protocol: layered messages with applied symmetric and asymmetric encryption, like we learned in class.

Relays design. Each relay has a pool of messages. When the number of messages inside the pool reaches a defined limit, the relay samples messages from the pool uniformly, and starts sending them in random order to the next destination (possibly another MixNode).

System design. We designed the system to be based on mixnet in order to break the correlation between sent messages. This way, further hiding the identity of the users. In addition, we aimed to prevent attackers that use passive network-surveillance and traffic analysis to correlate between messages entering and exiting the mix network. Furthermore, we want the MoT to get only

the basic information for its public transportation statistics, without knowing the identity of the users behind those rides.

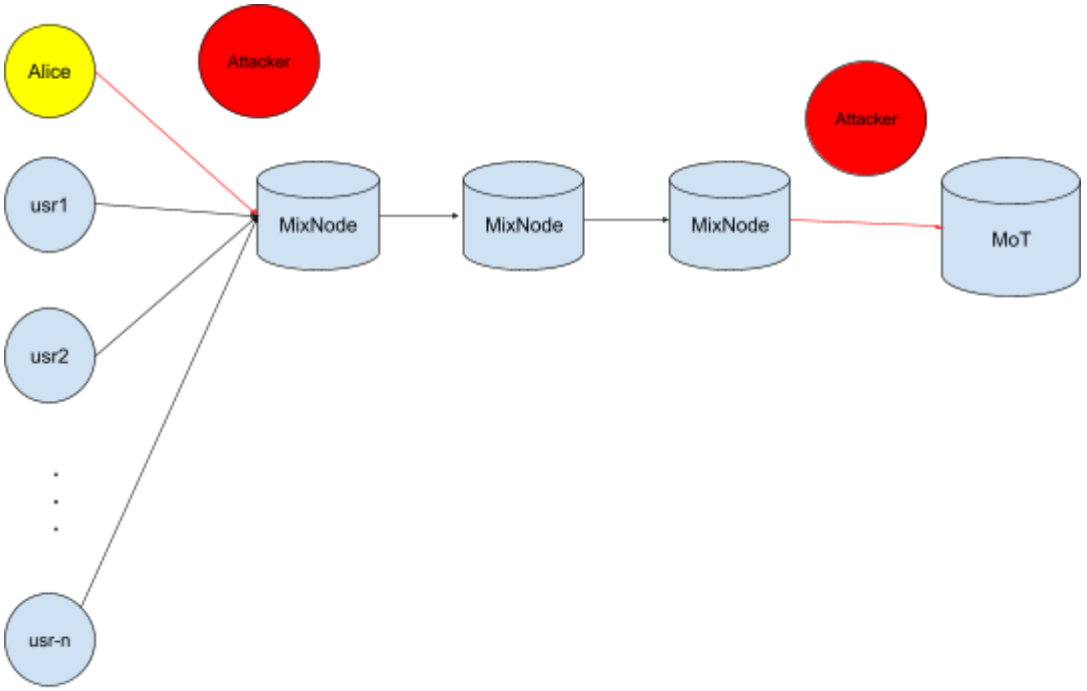


Figure: an illustration of workflow the system

3.5 Comparison to Existing systems

Today's applications systems, for instance, "Hopon", "Moovit", hold large amounts of user's sensitive data on their servers. According to their privacy policy, this includes: IP locations, general areas, cell phone id, Browser Information , Geographic location, Time of Visit ,date and time stamps of use, and all the rides information related to users.
We can deduce that these applications do not use anonymizing techniques like mixnet, or onion routing. All the private information of users they hold can be used against them legally (for police use for instance) or illegally. Furthermore, a breach in a system can cause personal information to be leaked outside.

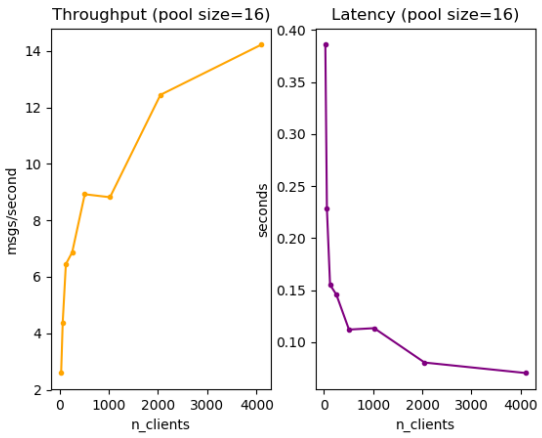
3.5.1 How does our system improve existing solutions?

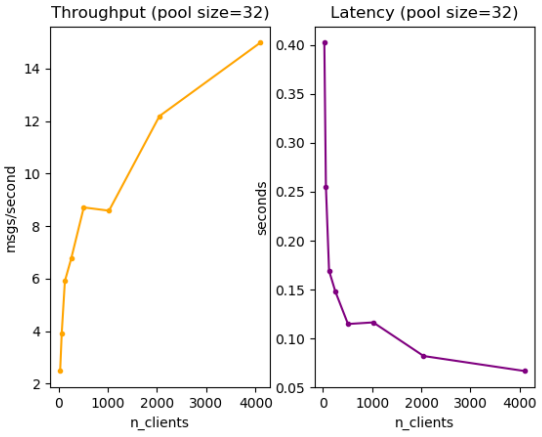
Our system breaks the correlation between users' rides data and personal information, and can only know what public transportation rides occurred, but cannot know by whom a ride has been made. The MoT is not exposed to sensitive and private information of users either.
These characteristics protect the users' privacy and sensitive information.
A passive attacker that watches the entering and exiting link of the mixnet, cannot deduce correlation between messages entering the mix network, and those that exist the mix network. Even if only one relay is honest, because the

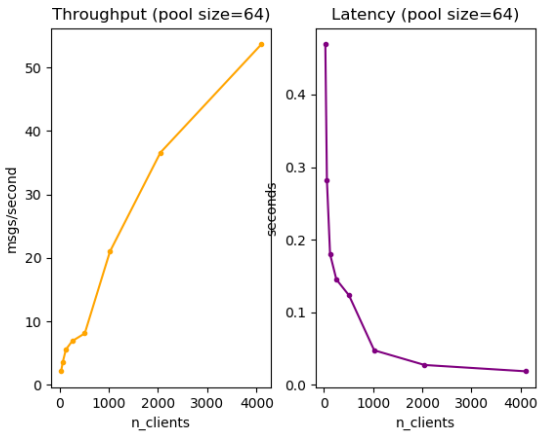
messages are sampled randomly, shuffled, and then being sent to the next hop, an attacker still cannot learn any information.

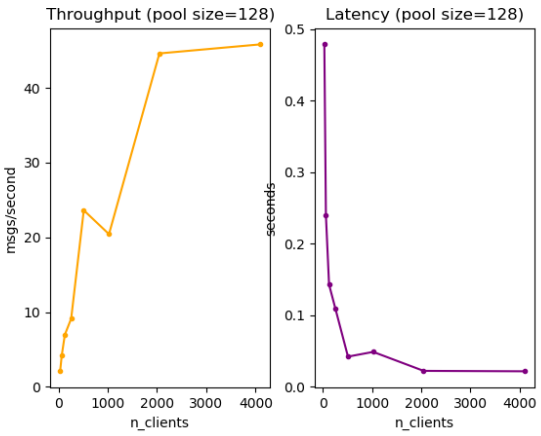
4 Evaluation

We measured the average latency and average throughput of the system. The evaluation was done on the local network only using 3 MixNodes/relays instances. Therefore, we cannot infer or estimate based on the evaluated performances regarding real-world usage.
All the network components were executed by a different thread on the computer system. specifically, The MoT-server, MixNodes, and the different clients. Each one of the clients sends 4 different messages to the mixnet, at different times.









Conclusion. We can see better average throughput performances with larger messages pool size limits: 64 or 128. But with no major changes with respect to the average latency. We chose a 64 pool size limit for the system.