Week 8 Blog

Sree Ram Boyapati	a1775690@student.adelaide.edu.au
Course	Comp Sci 7092 - Mobile and Wireless Systems
Group	CovidGuard-F
Supervisor	Zach Wang

What have I done this week?

In this previous week, I have done a bit of code cleaning and correcting a couple of bugs.

Issues from the demo implementation -

1. Advertising of older TEKs was not stopped when new TEKs were being generated. After running for a long time, a new RPI advertisement was failing.
   Ref: Error Code 2 in Beacon Transmitter
2. Synchronisation of Tek and RPI Generator was done using shared variables and it was not extensible. There is no tight bound of RPI within the timeframe of TEK. If the TEKs are downloaded, Few RPIs will never be matched because of it.
3. Previous Implementation of SQL based storage was not extensible and three databases were being created. Encryption of database was not straightforward.
4. Network calls happen in a background thread. However, Fetching of response was a blocking call. The screen used to get stuck waiting for the response.

To solve the three database problem, I have followed the Android Principles and made interaction with the database using ROOM[1] Library.
https://github.cs.adelaide.edu.au/2020-Mobile-and-Wireless-Systems/CovidGuard-F/pull/45

Benefits of using an ORM Layer are -

1. Verification of SQL queries at Compile Time
2. Integration with advanced android principles like RxJava3 and LiveData and SQLCipher for encryption.
3. No need to write raw SQL queries which can lead to
4. Export of schema seamlessly for verification and auditing.

[1] "Room Persistence Library | Android Developers." 27 Aug. 2020, https://developer.android.com/topic/libraries/architecture/room. Accessed 3 Oct. 2020.

To solve the Advertising Bug and Generation of RPI within a strict interval, I have refactored the flow of RPI Generation.
Ref:
https://github.cs.adelaide.edu.au/2020-Mobile-and-Wireless-Systems/CovidGuard-F/pull/46

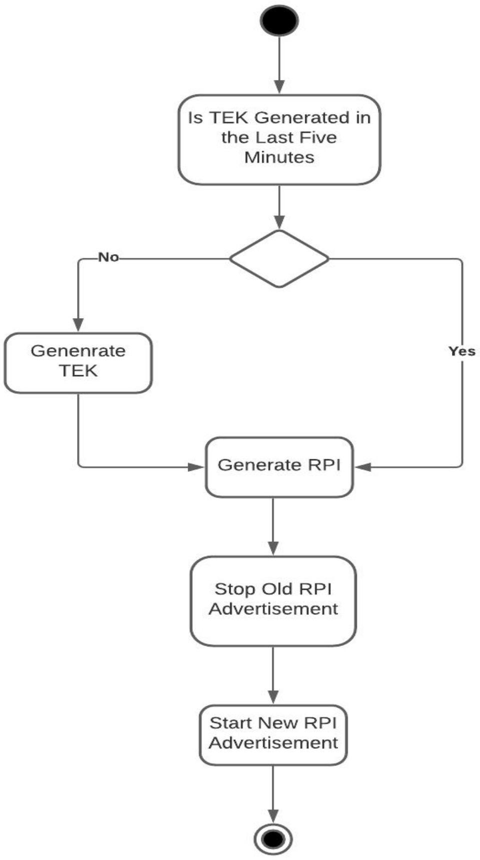

Is TEK Generated in the Last Five Minutes
No
Yes
Genenrate TEK
Generate RPI
Stop Old RPI Advertisement
Start New RPI Advertisement

Figure 1: New RPI Generation Service

Apart from this, I have created boilerplate code for Exposure Notification and Laboratory Server for separation of concerns. Exposure Notification Server will share teks of infected users and verification servers will store anonymous user details and Laboratory server will register with the user and knows the identity of the user. and verifies test results for the verification server. We have used covidWarn App[2] by germany for reference in our implementation.

https://github.cs.adelaide.edu.au/2020-Mobile-and-Wireless-Systems/CovidGuard-F/pull/47
https://github.cs.adelaide.edu.au/2020-Mobile-and-Wireless-Systems/CovidGuard-F/pull/42

[2] "Corona-Warn-App." https://www.coronawarn.app/. Accessed 3 Oct. 2020.

Other minor enhancements were -

1. Using Square's Retrofit[3] to remove a lot of boilerplate for making network calls. https://github.cs.adelaide.edu.au/2020-Mobile-and-Wireless-Systems/CovidGuard-F/pull/48
2. Using Facebook's Stetho[4] for profiling the app and database using google chrome.

[3] "Retrofit - Square Open Source." https://square.github.io/retrofit/. Accessed 3 Oct. 2020.

[4] "Stetho - Facebook Open Source." http://facebook.github.io/stetho/. Accessed 3 Oct. 2020.

Paper Review # 1

Koukis, Dimitris et al. "On the Privacy Risks of Publishing Anonymized IP Network Traces." *Communications and Multimedia Security* (2006).

Why have I chosen this paper?

Assessment of our solution is an integral part of the project and It is very important that the information we share as part of the system is scrubbed of private info like tokens or client ip addresses. We are presently receiving the client ip address as protoPayload.ip in the google cloud logs.

If we have to share network logs for intrusion detection or for checking against attacks in real time after post processing - Server logs may have to be shared with third-party vendors and auditors to validate for compliances adherence.

In our previous paper [1], We have studied a scalable version of tcpdpriv[5] that can anonymise IP addresses without taking the order of logs so large files can be parallely processed or logs can be streamed from all services to the central processor to anonymise and detect in real-time.

In that aspect, It is important to understand the kind of attacks on privacy that are possible with anonymised network traces.

Conclusion of the research

The author lists two kinds of privacy attacks from the network traces.

1. Re-construction of web APIs called or visited by identifying distinct web pages from a host of requests.
2. De-Anonymization of IP addresses from the address logs.

Together, User's browsing history can be determined. By grouping elements of requests that are done in order or periodically in a single connection, One can get the signature of a webpage. Fingerprint of a web page is a collection of signatures with some payload size differences taken into account due to variance in each user's data. By visiting the websites themselves and keeping track of fingerprints received from the traces collected by attackers can serve as validation data. Static websites were shown to have better tracking.

[5] "Program for Eliminating Confidential Information from Traces." http://fly.isti.cnr.it/software/tcpdpriv/. Accessed 5 Oct. 2020.

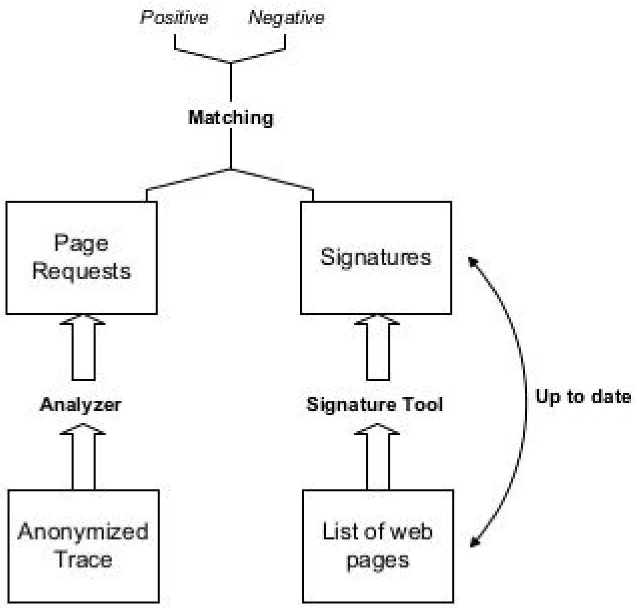


Figure 2: Fingerprinting Methodology

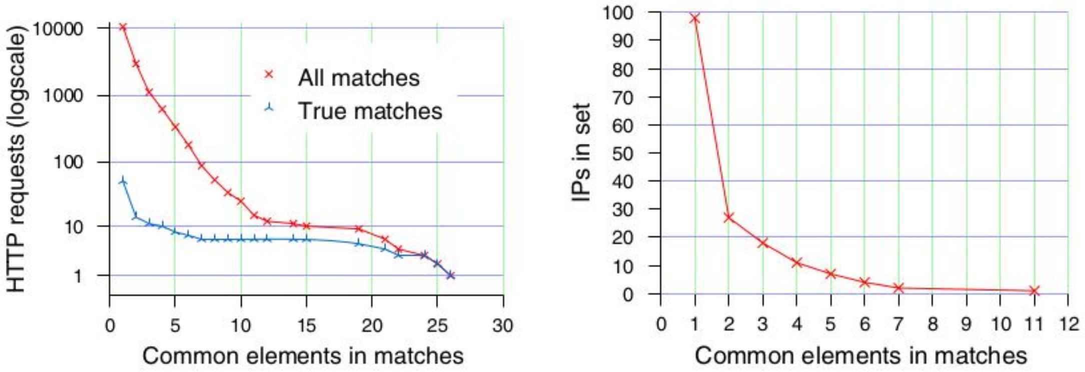
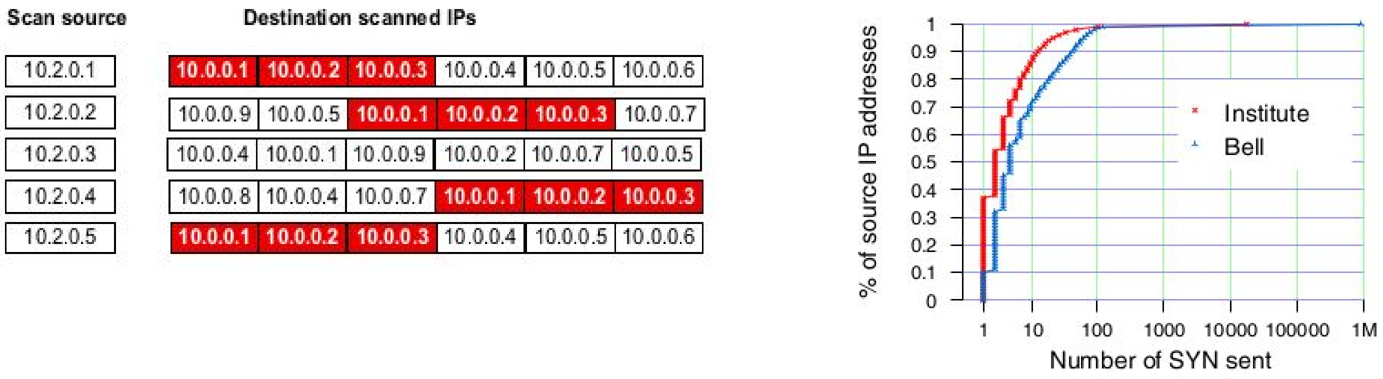


Figure 3: True vs All Matches w.r.t common elements in the dataset

Second part of the problem is deanonymization of traces. Instead of injecting scans and exposing oneself as an attacker, the authors propose to get existing scans. The authors depend on two assumption

1. Two scans that are identical may be sequential.
2. larger the common subsequence between different scans, probability of it being a linear scan is higher.
3. Common Subsequences can later be used to create the original ip addresses from the mapping used.

The authors ran their experiments on a smaller subnet of a.b.c.d/24. In the existing scans, Author identified the scanners by the number of SYN packets they have sent. They have discovered that few IP addresses sent most of the SYN packets. After getting the scanners, They tried to map the destination hosts of the scanners and check subsequence matching between different scans to get the info on real ip addresses.



Scan source	Destination scanned IPs					
10.2.0.1	10.0.0.1	10.0.0.2	10.0.0.3	10.0.0.4	10.0.0.5	10.0.0.6
10.2.0.2	10.0.0.9	10.0.0.5	10.0.0.1	10.0.0.2	10.0.0.3	10.0.0.7
10.2.0.3	10.0.0.4	10.0.0.1	10.0.0.9	10.0.0.2	10.0.0.7	10.0.0.5
10.2.0.4	10.0.0.8	10.0.0.4	10.0.0.7	10.0.0.1	10.0.0.2	10.0.0.3
10.2.0.5	10.0.0.1	10.0.0.2	10.0.0.3	10.0.0.4	10.0.0.5	10.0.0.6

Figure 4: Subsequences of matched IP addresses between different scans

Outcome of the research

1. Anonymized traces are not completely anonymous and the casual nature of the traces can be leveraged to build the context. https://slack.engineering/tracing-at-slack-thinking-in-causal-graphs/ - In fact, this is being leveraged by slack to identify bottlenecks in its microservices.
2. The paper was written in 2006 and the http/2 standard is not taken into account. Http/2 has provided multiplexing by default instead of HTTP pipelining. Further analysis is needed if we can extract the web signatures from elements in the same fashion.
3. The datasets analysed in the paper conclude that the web is majorly static as shown with true cases vs all matches in figure 4. In 2020, Web is more and more dynamic. Can Website fingerprinting methodology even take multiple signatures for each website and doing a threshold match for positive case be even relevant today. The precision was higher when the dataset was higher.
4. Most of the concerns arise from temporal properties of the logs. Request A is followed by Request B. To share with third-party vendors, can only the shuffled network trace be given at the loss of few security features that can be leveraged from temporal data?
5. Deanonymization of IP addresses is still possible, however the direct mapping from subsequences of anonymised traces to real IPs has not been discussed in the paper. In [1], Some semantic attacks using frequency analysis have been described to get the mapping.

References

1. Jun Xu, Jinliang Fan, M. H. Ammar and S. B. Moon, "Prefix-preserving IP address anonymization: measurement-based security evaluation and a new cryptography-based scheme," 10th IEEE International Conference on Network Protocols, 2002. Proceedings., Paris, France, 2002, pp. 280-289, doi: 10.1109/ICNP.2002.1181415.