

# Traffic Analysis Threats to Private Web Browsing

George Danezis



## 1 Description of the Problem

- **SSL mainly protects the content** of a communication by applying encryption techniques to guarantee its confidentiality and integrity.
- **SSL does not reveal the exact path** of the resource accessed. But the network address of the Web Server is known to the attacker. This already reduces the number of candidate documents that a user might have accessed.
- **SSL does not hide the length** of the resource accessed, although some blurring occurs because of padding. Additionally more than one documents could be transferred during the same connection, making it difficult for the attacker to detect their boundaries.

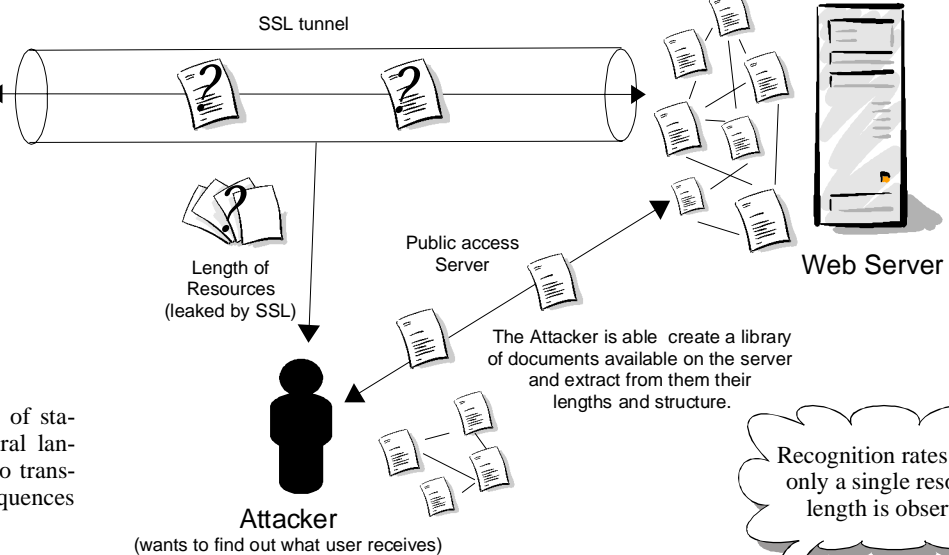


## 2 Privacy & the Web

- **The “Big Browser” threat:** Third parties being able to reconstruct exact browsing patterns of users. Became a subject of great concern during the RIP Bill discussions in the UK.
- These patterns of user behavior can be used to **identify users**. Identification can happen because the same user will tend to visit a particular set of pages regularly.
- The browsing patterns can also be used to **extract personal information**. Even more information could be extracted since many users are not aware that their browsing could in fact not be private.
- **Profiles** could also be designed and tested against particular user behavior on the Web.

## 3 Traffic Analysis

- Historically traffic analysis is used in signal and communication intelligence by the military. They were developed shortly after wireless communications were invented, and their main objective is to identify and locate enemy units and extract information about their status and intentions.
- Attackers try to **guess which pages a user is accessing** using the sequence of lengths an attacker can see on the encrypted communication link.
- We assume that the web server is public, so that the **attacker can create a good model** of how users would navigate through it. The attacker also knows the length of resources on the server.

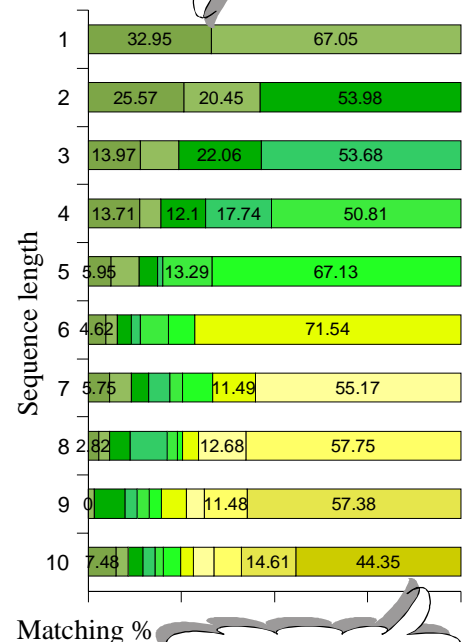


## 4 Techniques

- Borrow techniques from the fields of statistics, machine learning and natural language processing, and adapt them to transform sequences of lengths into sequences of resources.
- **Match the length observed on the encrypted channel with length of documents** on the server. This might lead to both false positives and false negatives, since some resources have the same length.
- **Model the navigation behavior of users to disambiguate between resources with the same length.** Server logs, or its static structure can be used to model the usual paths that users take through the system.
- The above techniques require the a-priori knowledge of the lengths of the resources on the Server, and the structure of the Web pages. They also require the attacker to observe the SSL packets on the communication channel. Aggregate lengths of communications or poor knowledge of the resources will seriously reduce their effectiveness.

## 5 Results

- **In most cases the attacker can determine correctly which resources users are accessing.**
- As the graph shows, **analysing longer sequences of requests provides much better results.** When analysing longer sequences partial matches are also found and can be considered as a success depending on the application.
- The recognition engine **can handle quite well fuzzy or incomplete data** and performs well even when some of the lengths are missing.
- Limits: The attacker can only determine links between users and resources, but **cannot guess any fixed length content.**



## More Information & Contacts

George Danezis  
 Computer Laboratory, JJ Thomson Avenue, Cambridge CB3 0FD, U.K.  
 Phone +44 1223 7-63569  
 Email [George.Danezis@cl.cam.ac.uk](mailto:George.Danezis@cl.cam.ac.uk) -Web <http://www.cl.cam.ac.uk/~gd216/>