Project Report Template

Sayed Hadi Hashemi, Roy H. Campbell
University of Illinois at Urbana-Champaign

1 Introduction

To improve product quality and user experiences, more and more companies and collecting user's anonymous data []. Typically, it will be three steps for collecting anonymous data []. First, gather user's insensitive information on client side; Second, send the information back to companies database or other cloud platform []. Third, analyze collected data. For each of step, there is possibility of leaking user's sensitive data. In the first step, the client side software or JavaScript may collect user's privacy actively, which is known as malware or spyware. In the second step, user may expose sensitive information such as ISP or Geo-location due to the network transmission which can be passively collected. In the third step, the companies may find correlation between the anonymous data they collected with other open data to review user's identity. To protect people from leaking their sensitive information, we have software analysis to deal with malware and spyware in the first stage [] and differentiate privacy [] in the third stage. However, we do not have many solutions to prevent data leak during the second stage.

Sensitive information can be easily leaked in the second stage, via device fingerprints [], even Tor solve the problem []. Typically, device fingerprints can be acquired passively through browser's behavior or even lower layers of TCP/IP model. When the user connects directly to the company's server to submit anonymous data, the user's public IP address can easily be acquired in transport layer and the physical location will be exposed by querying IP Geo-location database []. Considering most companies are using HTTP or HTTPS protocol for data transmission, they can easily acquire more sensitive data by checking HTTP header. For example, ”X-Forwarded-For” field tells the company if the user is using proxy and what the original IP is, and ”User-Agent” field tells which browser and operating system the user is using. When it comes to web applications, there will be more ways to track user down. The web application provider can use DNS Leak [] to find out which DNS server the user is using and use WebRTC [] to acquire the user's internal IP address. Besides, the JavaScript's behavior, screen resolution, Flash and Java plugins are other side-channels can be used by the companies []. According to Mozilla's report, ”83.6

Traditional ways of privacy protection in second phase are using VPN or Tor to hide the user's real identity. However, these techniques which are trying to hide user's IP are not enough. By using device fingerprint, the company can still track the user behind the Tor []. To solve the problem, we introduce our P2P based anonymous data collection system. Instead of trying to hide the identity, we decouple the relation between device fingerprint and collected data. Tracking the origin of anonymous data will be liking finding needle in a haystack.

As shown in figure ??.

2 Related Work