Fingerprinting Past the Front Page: Identifying Keywords in Search Queries over Tor
Se Eun Oh                       Nicholas Hopper
University of Minnesota

Abstract
• In this work, we introduce a Keyword Fingerprinting (KF), extending Website Fingerprinting (WF), to identify keywords in search queries. Based on a two-stage, traffic analysis-based approach with new task-specific feature sets, a passive network adversary can defeat the use of Tor.
• We demonstrate the feasibility of the KF attacks across four popular search engines and various experimental settings (e.g., user query setting). We also further explore why several keywords are better fingerprintable.

Keyword Fingerprinting (KF)
• The attacker will progress through two sequential fingerprinting steps.
  • 1st step: Webpage fingerprinting to identify the query result traffic of the specific engine
  • 2nd step: KF to predict keywords in query traces by both binary and multi-class classification
• KF focuses on 2nd step, which is challenging for existing WF techniques.

KF vs. WF
• CUMUL classifiers proposed by Panchenko et al. perform very well for the 1st step, which detects blue against green area. However, when identifying and differentiating keywords in blue, classifiers based on WF features perform poorly.

RESP feature set
• All 80,000 query traces included a long sequence of incoming packets at the end of the trace. We call it “RESP” and remaining portion “Request”
• RESP is more informative than the request portion

Data Preparation
• Reverse cumulRespTLS and cumulRespToCell
• The last elements are total size of TLS records and total number of Tor cells in RESP and good features to identify search terms
• SVM accuracy for the first and last 140 packets in cumulRespTLS: 21.33% vs. 53.79%
• Number of Features: Use 247 features as it gave the best accuracy as well as acceptable running time

Feature evaluation using $\chi^2$ statistics
• We tested different combinations of feature sets whose $\chi^2$ statistics was higher than 600, and the best feature set was “Agg” aggregating Total, RespTotal, RcumulRespTLS, and RcumulRespToCell

Support Vector Machine
• We used a non-linear classifier with a radial basis function (RBF) and 10-fold cross validation to find $C^*$ and to split dataset into training and testing set.
• Metrics
  - Binary Classification: Precision, Recall (TPR, FPR) (%)
  - Multi-class classification: Within-monitored Accuracy (WM-acq) (%)

TPR and FPR when we identify 10k Google and Duckduckgo query traces against 100k webpage traces
• Google query trace identification
  - Ratio
    - TPR (%): 99.94, 99.94, 99.86, 99.84
    - FPR (%) 0, 0, 0, 0
    - precision (%) 100, 100, 100, 100
  - Duckduckgo query trace identification
    - Ratio
      - FPR (%) 0, 0, 0, 0
      - precision (%) 100, 100, 100, 100

Closed and Open World Experiment
• Closed-world accuracy (10k keywords and 100 classes)
• Identifying 100 monitored keywords against 10k background keywords
• Identifying 100 monitored keywords against 10k background keywords

TPR and Analysis on search result HTML
• Effect of Label Learning (Binary vs. Multi)
• Effect of Search Engines (Google vs. Bing vs. Yahoo)
• Effect of Query Setting (US enabled vs. disabled)