Efficient Large-scale Collection of Information Related to Domain Names

(for the purpose of malicious domain name detection)

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Author: Ing. Ondřej Ondryáš
Supervisor: Ing. Radek Hranický, Ph.D.

Motivation

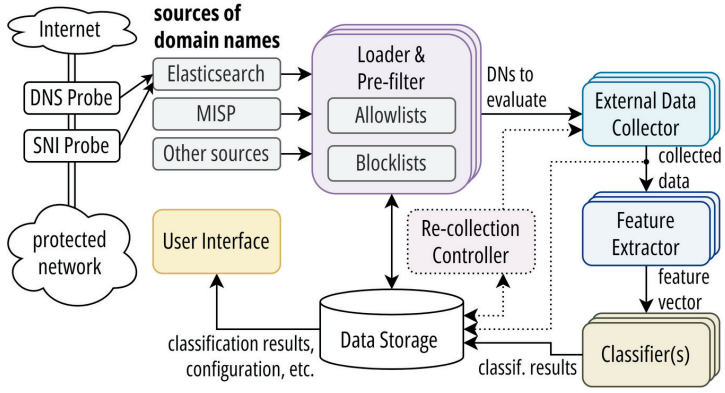
The internet is a seedbed of malicious and criminal behaviour targeted both at individuals and institutions. Some attackers employ **phishing**, tricking someone into giving their credentials directly to them, for example, by convincingly imitating a legitimate service. Others attempt to install malicious software – **malware** – on the users' devices. This software may track the user, send their data to the attacker or convert the device into a puppet following the attacker's commands.

While satisfactory results have been demonstrated with methods that detect malicious traffic from the structure of web pages or URLs, their usability is seriously limited by the prevalence of **encrypted communication** and so techniques independent on the contents of the traffic are needed. One promising approach is based on the analysis of **domain names**. When accessing online services, clients resolve the names mostly through unencrypted DNS [1]. They can be captured at the perimeter of a protected network and evaluated using ML classifiers.

On their own, domain names do not provide many threat indicators. However, enriching them with external information, e.g. using DNS scans or RDAP/WHOIS to gather registration data, significantly enhances the threat assessment [2]. A robust and scalable data collecting and processing solution is necessary for real-time monitoring of domain names in large networks.

Context and Goals

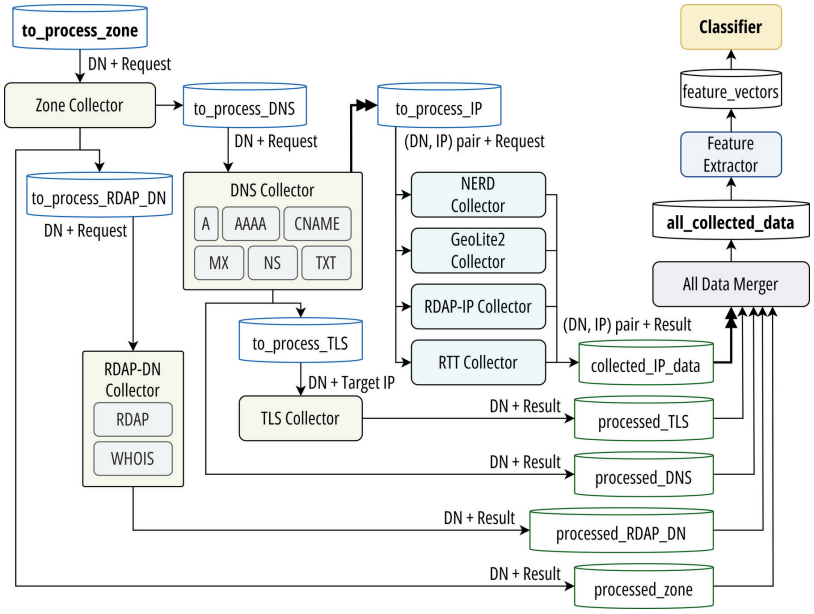
The thesis introduces a data collection pipeline that forms the basis of the **DomainRadar** system. It accepts domain names from various sources. For each, it determines the **authoritative namesevers**, performs a **DNS scan**, collects registration data through **RDAP or WHOIS**, does a **TLS handshake** in order to store certificate data and protocol configuration, and for each **IP address**, it collects **geolocation** data, **autonomous system** affiliation, **registration** data and **reputation** data from CESNET's NERD. The collected data are processed by the feature extractor to compute **147 features** adopted from literature or designed by the author and other members of the DomainRadar team.



A high-level scheme of the DomainRadar system.

Design

The system is designed as a set of microservices that communicate through the **Apache Kafka** platform. *Collectors* are responsible for gathering data from the external sources and producing requests for the following stages of the pipeline. The collected data are merged and processed by the *feature extractor* and, ultimately, the classifiers. The results of the collection and the classification are stored in database systems using **Kafka Connect**. The use of Kafka provides scalability: each Kafka topic (that represents a connection between two components) can be partitioned – Kafka then splits the processing requests evenly between multiple instance of a component. Scaling the system horizontally is as simple as adjusting Kafka's configuration and executing more instances of the components.



Evaluation and Results

The system was tested on large static set of domains and using capped real-time traffic. It processed:

- 400,000 domain names in 4 h 16 min, resulting in average throughput of 22.65 DN/s
- 9.54 DN/s of real traffic from the CESNET network (limited by the input)

CPU and memory usage were monitored. In the first case, the deployment saturated the 4 vCPUs @ 2.9 GHz and 16 GiB of RAM. In the second case, the system was underutilised at about 18% CPU usage and 11.83 GiB of RAM on average.

[1] K. Hynek, "The Impact of Encrypted DNS on Network Security," Ph.D. dissertation, Faculty of Information Technology, Czech Technical University in Prague, Prague, 2023.

[2] R. Hranický, A. Horák, J. Polišenský, K. Jeřábek and O. Ryšavý, "Unmasking the Phishermen: Phishing Domain Detection with Machine Learning and Multi-Source Intelligence," NOMS 2024-2024 IEEE Network Operations and Management Symposium, Seoul, Korea, 2024.