Identifying URLs for Blacklist

Goals
The following system is designed to enable the detection of malicious URLs that should be blacklisted.

In recent years, many attacks originate from surfing a malicious page on the Internet. Creating such a “black list” of malicious URLs is a goal of many companies in the industry. Surfing to a malicious URL can cause harm to the specific user and in many cases to the whole network to which the computer belongs. Categorizing a URL as malicious is not trivial for many reasons. First, in many cases the page is hiding behind a short URL that is created by some URL shortener. Second, the page looks normal but it causes the system to perform abnormally. Lastly, in many cases, the behavior of the system looks normal and, by looking at a single user surfing to the malicious page, the URL cannot be detected.

Description
We first need to create a normal profile of the system. This can be done by logging the system behavior when there is no Internet connection. This stage should be done on many PCs so we can correlate and filter the normal behavior for the next stage.

In the next stage we will run a crawler; we log each page the crawler is entering and the behavior of the system during the downloading time. This is, again, done on many PCs.

The third stage should be to filter the normal behavior from the log of the second stage and to identify pages that cause abnormal behavior such as writing to unwanted places. The output of this page is a list of suspicious URLs.

The last stage would be to analyze these suspicious URLs by entering them again, but this time with a deeper analysis such as logging the written files.

The goal of this project is to be able to identify any malicious URL with zero false positives, i.e. to identify only the malicious URLs. By having a complete and accurate list, attacks that enter the system through malicious URLs can be stopped.

Many companies in the industry maintain black list of URLs. Still, none of them have a complete list. Such a list can be embedded in any Intrusion Prevention System (IPS), or even in the core of the network.