Web site defacement, the process of introducing unauthorized changes to a web site, is one of the most common attacks on the Internet. More than 390,000 web sites have been defaced in 2004 and the trend has been constantly growing in the recent years (see web server intrusion statistics on http://zone-h.org/news/read/id=4457, for example). A defacement consists of a fraudulent replacement of most or all of the content of a site with other content aimed at making the intrusion evident to a visitor. Obviously, an attacker capable of performing a defacement would be equally capable of introducing subtler changes difficult, if not impossible, for a user to point out — e.g., modifying a script that collects username and password sent by the user, so that these credentials are also sent to a location chosen by the attacker. There is clearly an urgent need to develop methodologies for effectively addressing this issue.

The only existing approach to automatic detection of such attacks is based on a comparison between the web resource and a copy of the resource kept in a safe place and assumed to be uncorrupted. Implementing such a framework may be expensive and difficult: there is a plenty of web sites from which many people depend upon, yet most of the organizations hosting these sites have neither the expertise nor the budget necessary to buy, install, configure and maintain one of the existing, expensive tools or services based on the “trusted version” approach.

We explore a radically different approach and propose a tool capable of monitoring the integrity of remote web resources automatically, without any form of involvement from the site being monitored and, in particular, without requiring the installation of any infrastructure at that site. The tool fetches the monitored resource from the web at regular intervals, analyzes it, and generates an alert whenever it detects that a form of “unusual” change in the resource has occurred. The tool only needs an initial observation period for constructing an internal model of the resource.

We evaluated our tool as follows. We selected 6 sites totalling 11 different pages. The sites were chosen to represent a suitable mix of real-world sites including, in particular, highly dynamic contents: CNN (home page, business main page and weather main page), Amazon (home page), Autoroutes de Sud de la France (home page and traffic page), ANSA (home page and main RSS feed), Repubblica (home page and technology and science main page). Each page has been observed for 15 days and then monitored for 20 days. We fetched a copy every 6 hours, thus building a test set of 80 snapshots for each page. In the monitoring period, we first counted the number of false positives raised by the tool. Then, we simulated 20 different defacements and counted the number of missed detections, i.e. the number of false negatives. In the attempt to mimic “real world defacements”, we extracted randomly 20 mirrored defacements from the “Digital Attacks Archive” at www.zone-h.org.

The tool exhibited neither false negatives (i.e., all defacements were detected) nor false positives (i.e., the tool did not raise an alert in spite of the dynamically changing content). These results are important since they show the ability of our approach to cope with dynamic content and discriminate between legitimate and unauthorized modifications, while not requiring any prior knowledge of the monitored resource (except, of course, from the initial observation period). This indicates that our approach may be indeed viable in practice.

This framework may allow developing services capable of monitoring the integrity of many foreign web sites cheaply, which may be very attractive for small business and administrative organizations that depend on the web for their operations but have tight budget constraints. Indeed, this has been the main driving motivation for our work.