In this paper, author introduced us a new approach to predict that if a website will become a malicious one or not within a period of time, which implemented with machine learning, data mining.

The prediction first begins with building a classifier by learning the characters of both malicious websites and benign ones, with extracting features from data received from Internet. For example, by randomly sampling DNS zone files to collect raw data. To make sure the approach is effective, could be able to predict the website is going to become a vulnerable one, the author decides to use past archived versions of the site to collect feature data, since you have already known the status of website now, then it would be possible to test the result of prediction is correct or not(page 628).

The approach selected several features to evaluate the status of a website, including traffic statistics, filesystem structure, Webpage structure and contents.

Then in the demonstration, the paper showed how this prediction approach working in practice(page 6). After data has been collected from existing black list or normal network, author picked one of the feature, the number of pages in a website to extinguish websites and evaluate they are threat or not. And also, choose the early version within the time horizon of malicious websites, then extract the webpage template with a parameter called “composite importance” (page 631), and according the test result, seems scraping 20 pages for each website is a reasonable value

With the template extracted from the websites, then author evaluated them with features primarily focused on AWIS and Content-based. With the features selected by feature extracted by static and dynamic feature extraction, then the system will be able to figure out If the website is an malicious one(page 632)

After reading this paper, the author convinced me that this approach is effective to be used in predicting the website is compromised or not, with data mining serving the system to learn about features that a malicious website can probably have, this approach can probably make the right prediction.

However, on the other hand, the approach also has some limitation. As discussed in the paper, sometimes the website can possibly be compromised not by its characters, by only because of other security issue, such as default password or database positon. If the approach that features of this kind of website into account, the result of calculation will definitely be compromised. What’s more, the threat that this approach primarily focused on seems to be changing the content or structure of pages dramatically for advertising or redirecting propose, so I’m a wondering how could this system detect those websites that being compromised but with no changes on its website or only a little bit. Maybe the attacker just wants the hardware resources of the website server. And as for me, it’s not very easy to understand what this methods core with all these mathematic expressions.

In my opinion, the team of this paper may consider putting their approach into a portable application for websites in different environments as further investigation. Before achieving this goal, maybe then just need to improve the compatibility for their users.