# WhoisXML API data supports a comprehensive study of deceptive internationalized domain names

Posted on August 10, 2020

Cybersquatting is a kind of threat that is, unfortunately, becoming more and more prevalent. Cyber attackers register a tremendous number of domains every day that are confusable with those of legitimate and popular brands. These domains are then used to trick victims into thinking that they are actually visiting a legitimate page. A common approach is typosquatting, a slight misspelling of the domain name, e.g., writing "faceboek" instead of "facebook."

This problem aggravated with the system of international domain names (IDNs)., Indeed, the possibility of using any Unicode character in a domain name was standardized and implemented in 2003. Unfortunately, that introduced a new squatting strategy: the use of non-English national characters to confuse victims. This approach has become increasingly popular amongst miscreants.

Daiki Chiba of NTT Secure Platform Laboratories, Tokyo, Japan, and his co-authors, in collaboration with Waseda University, Tokyo, Japan, have recently published, perhaps, the most comprehensive study on this kind of attacks. The study was presented during the 22nd International Symposium on Research in Attacks, Intrusions and Defenses, organized by the USENIX Association and held in Beijing in September 2019.

The study covers domain names targeting both English and non-English brand names. They investigate three types of domain names. In homo IDNs a few characters are replaced with similarly looking international ones, like "fácebook[.]test". Combo IDNs add some extra string (like "example?? ??[.]test"; the Japanese characters stand for "login") to the original domain name. The two techniques are also used in combination by the miscreants, leading to "homocombo" IDNs.

As part of their work, the researchers propose a new system called DOMAIN SCOUTER to automatically detect the six types of deceptive IDNs (eng-combo, eng-homo, eng-homocombo, noneng-combo, noneng-homo, and noneng-homocombo). With a set of domain names and a list of (both English and non-English) brand names, the system can detect the deceptive domain names that are likely to be registered for malicious purposes. A "deceptive IDN score" is assigned to the domain names to quantify this. That is achieved by processing various features focusing on visual similarities, brand information, and TLD characteristics.

To accomplish their goal, the authors needed a complete list of all registered domains. They opted for the WHOIS database provided by WhoisXML API, possibly the most comprehensive list of this kind on the market. They processed over 294 million domains (including IDNs and non-IDNs) under 1,435 TLDs. From all the domains, they extracted over 4.4 million IDNs under 570 TLDs

(the others did not contain any of this kind). This data set formed a solid basis for the study. The detection method made use of additional features, notably, the creation dates of the domains, which are also available in the WHOIS database.

Beside convincingly demonstrating that their automated detection technique is indeed powerful, they also conducted online surveys that highlight how deceptive IDN domains can unfortunately really confuse many potential victims. The work concludes with suggestions for client applications, domain registrars/registries, domain owners, and certificate authorities to reduce the risk of deceptive IDNs.