The Homograph Attack

Evgeniy Gabrilovich   Alex Gontmakher
gabr@acm.org   gsasha@cs.technion.ac.il

Computing veterans remember an old habit of crossing zeros (Ø) in program listings to avoid confusing them with the letter O, in order to make sure the operator would type the program correctly into the computer. This habit, once necessary, has long been rendered obsolete by the increased availability of editing tools. However, the underlying problem of character resemblance is still there. Today it seems we may have to acquire a similar habit, this time to address an issue much more threatening than mere typos: security.

Let us begin with a short recourse to history. On April 7, 2000 an anonymous site published a bogus story intimating that the company *PairGain Technologies* (NASDAQ:PAIR) was about to be acquired for approximately twice its market value. The site employed the look and feel of the *Bloomberg* news service, and thus appeared quite authentic to unsuspecting users. To disseminate the "news", a message containing a link to the story was simultaneously posted to the Yahoo message board dedicated to PairGain. The link referred to the phony site by its numerical IP address rather than by name, and thus obscured its true identity. Many readers were convinced by the Bloomberg look and feel, and accepted the story at face value despite its suspicious address. As a result, PairGain stock first jumped 31%, and then fell drastically, incurring severe losses to investors.

Attacks like this are relatively easy to detect. A stronger variant of this hoax might have used a domain named bl00mberg.com[1] (with zeros replacing o's), but even the latter is easily distinguishable from the real thing. However, forthcoming Internet technologies have the potential to make such attacks much more elusive and devastating.

A new initiative, promoted by a number of Internet standards bodies including IETF and IANA, allows one to register domain names in national alphabets. This way, for example, Russian news site "gazeta.ru" ("gazeta" means "newspaper" in Russian) might register a more appealing "газета.ру". Far from buzzword compliance, the initiative caters to the genuine needs of non-English-speaking Internet users[2], who currently find it difficult to access Web sites otherwise. Several alternative implementations are currently being considered, and we can expect the standardization process to be completed soon.

The benefits of this initiative are indisputable. Yet the very idea of such an infrastructure is compromised by the peculiarities of world alphabets. Revisiting our newspaper example, one can observe that Russian letters "а,е,р,у" are indistinguishable in writing from their English counterparts. Some of the letters (such as "а") are close etymologically, while others look similar by sheer coincidence. For instance, Russian letter "р" is actually pronounced like "r", but the glyphs of the two letters are identical. As it happens, Russian is not the only such language; other Cyrillic languages may cause similar collisions.

With the proposed infrastructure in place, numerous English domain names may be *homographed* – maliciously misspelled by substitution of non-Latin letters. For example, the Bloomberg attack could have been crafted much more skillfully, by registering a domain name bloomberg.com, where the letters "o" and/or "e" have been faked with Russian substitutes. Without adequate safety mechanisms, this scheme can easily mislead even the most cautious reader.

[1] Incidentally, this domain has actually been registered.

[2] According to Global Reach's report, the English-speaking population of the Internet was about 62% in 1998, and is forecasted to be as low as 37% by the end of 2002.

Sounds frightening? Here is something more scary.

One day John Hacker similarly imitates the name of your bank's Web site. He then uses the newly registered domain to install an eavesdropping proxy, which transparently routes all the incoming traffic to the real site. To make the bank's customers go through his site, John H. hacks several prominent portals which link to the bank, substituting the bogus address for the original one. And now John H. has access to unending streams of passwords to bank accounts. Note that this plot can be in service for years, while customers unfortunate enough to have bookmarked the new link might use it forever. And since most URLs today are clicked rather than typed, the outlook becomes quite chilling.

Apparently, the designers of the multilingual domains infrastructure have been aware of some dangers pertinent to the use of Unicode characters. Particularly, they forbid the use of many auxiliary characters that may cause confusion (for instance, the numerous variants of the "-" sign). In our case, however, the characters used for the fraud are perfectly legitimate.

Several approaches can be employed to guard against this kind of attack. The simplest fix would indiscriminately prohibit domain names that mix letters from different alphabets, but this will block certainly useful names like "`CNNenEspañol.com`". More practically, the browser can highlight international letters present in domain names with a distinct color, although many users may find this technique overly intrusive. A more user-friendly browser may only highlight truly suspicious names, such as ones that mix letters within a single word. For additional security, the browser can use a map of identical letters to search for collisions between the requested domain and similarly written registered ones. If necessary, it would then warn the user of suspected fraud.

Caveat. To demonstrate the feasibility of the described attack, we registered a homographed domain name `http://www.miсrоsоft.com`[3], which incorporates Russian letters 'с' and 'о'. While it may be tricky to type in, especially if your computer does not feature a localized keyboard layout, you can access this URL from `http://www.cs.technion.ac.il/~gabr/papers/homograph.html`[4].

So, next time you see "`microsoft.com`", where does it want to go today?

About the authors

Evgeniy Gabrilovich is a Ph.D. student in Computer Science at the Technion – Israel Institute of Technology. He is a member of the ACM and the IEEE. His interests involve computational linguistics, information retrieval, and machine learning. He can be contacted at gabr@acm.org.

Alex Gontmakher is a Ph.D. student in Computer Science at the Technion – Israel Institute of Technology. His interests include parallel algorithms and constructed languages. He can be reached at gsasha@cs.technion.ac.il.

[3] Predictably, `micr0s0ft.com`, `micr0soft.com` and `micros0ft.com` are all registered. John H. has not been wasting his time.

[4] Note that most browsers currently need a special client application *iClient* distributed by `i-DNS.net` in order to handle multilingual domain names. Also, some browsers might display this name in a garbled way (encoded in the ASCII/English version of the international characters as "`bq--at7w373jih7xepx7om7p6zx7oq.com`"). Naturally, when the infrastructure implementation is finalized, the name will be displayed correctly.