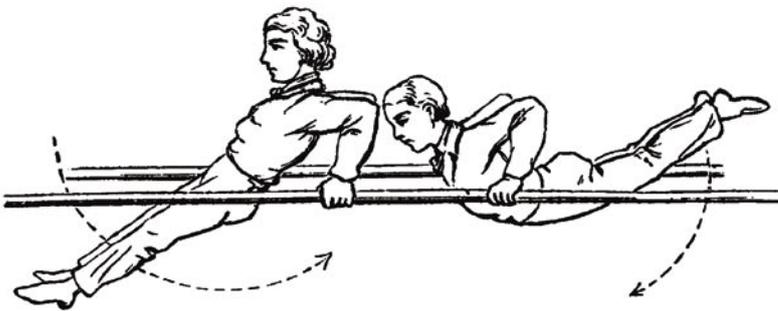# haking

Tricks of the Spammer's Trade

John Graham-Cumming

**English version of an article
published in "Hakin9" Magazine
issue 3/2004**

Tricks of the Spammer's Trade

John Graham-Cumming

Spammers try to get their messages through spam filters by using trickery. Let's see how these tricks work, and how up to date filters spot the trickery and use it to their advantage.

Spammers use three major types of trickery. They try to obfuscate bad words (such as *Viagra*) so that the filter doesn't see them, they try to include good or innocent words that the filter sees but the reader doesn't so that the filter thinks the message is legtimate and, knowing that the URLs of their web sites are often a giveaway, they try to obscure any URLs used.

Obscuring Bad Words

The first thing spammers try to do is hide the words that word give their message away as spam. They use a collection of techniques all designed to take a word like *Viagra* and render it unintelligible to a spam filter, but perfectly readable to a human.

Lost in Space
The simplest trick of all is to take a suspicious word like *Viagra* and space it out.

```
V I A G R A
```

This fools simplistic spam filters that search for the word *Viagra*; naturally a more sophisticated

spam filter can look for the pattern `<letter><space><letter><space>...` and reconstruct the suspicious word. Because of that spammers use a variety of other characters to space out the word:

```
V'I'A'G'R'A
V.I.A.G.R.A
V*I*A*G*R*A
V-I-A-G-R-A
```

And the list could go on and on. Unfortunately for the spammer it's pretty easy for a spam filter to look for these different patterns and figure out that the email is talking about Viagra. But this simple technique does raise a flag for anyone considering buying a spam filter: don't buy

---

What should you know...

- bayesian and heuristic filtering basics (see previous issue of our magazine),
- HTML and Javascript basics.

What will you learn...

- what tricks spammers use to go around bayesian and heuristic filters.

---

one that you are required to update with the latest rules; get one with an automatic update service. Even staying current with this simple way of obscuring a word would require a large effort on your part.

Spammers, of course, test their spams against free and commercial anti-spam software and have obviously realized that this specific trick isn't working well, and so they've moved on to changing the actual letters of *Viagra*. One spam I saw went the opposite direction and removed all spacing from the message and replaced the spaces with random letters:

```
DidAyouFknowNyouMcanBget
VprescriptionVmedications
prescribedTonlineTwith
NORPRIORRPRESCRIPTIONRREQUIRED!
```

This doesn't seem like a very effective technique, it makes the message almost unreadable by a human.

Ze Foreign Accent
A quick look at the ASCII table will reveal the presence of lots of accented vowels which spammers can use to take a suspicious word and obscure it by changing the vowels for their accented equivalents:

- a: à á â ã ä å,
- e: è é ê ë,
- i: ì í î ï,
- o: ò ó ô õ ö,
- u: ù ú û ü.

Just mixing and matching different accented a's and i's gives a spammer 144 different ways to write *Viagra*, such as *Víagra*, *Viågra*, *Vîãgrä*. English speakers just ignore the vowels and read the word, but a spam filter can be fooled.

Of course a spam filter programmed to recognize spammer trickery can map each accented vowel back to the basic letter to reconstruct the actual word. Since both this trick and the preceding one are easy pickings for today's spam

filters, spammers have turned to HTML for inventive ways to end up in your inbox.

A Numbers Game
Another way to hide the word Viagra is to use a special feature of HTML designed for inserting special or non-English characters. These HTML entities are written starting with `&#` and ending with a `;`. For example, to write the French accented character *é* in HTML you write `é`, to write the Greek letter ∑ you write `Ε`.

In fact all characters, including the standard English alphabet, have equivalent entities. The letter *A* for example can also be written `A` and so a crafty spammer can rewrite the entire word *Viagra* in entities: `Viagra`

Once again an up to date spam filter will understand HTML entities and do the conversion back to the real word. Much more sophisticated obfuscations of the word *Viagra* are possible by delving into HTML's formatting features.

Hypertextus Interruptus
HTML information about formatting is specified using what are known as HTML tags: instructions written between `<` `>` brackets.

For example, to take the word *Hello* and specify that it should appear in bold text you write: `<b>Hello</b>`. The `<b>` means *start bold text*, and the `</b>` means *finish bold text*. The text between the two tags will appear bold when displayed using a web browser, or an email program that understands HTML.

Like most computer languages HTML also has a mechanism by which the creator of a page or message can insert a comment. These comments are there for other people to read, but are completely ignored when the HTML is displayed. A comment starts with `<!--` and ends with `-->`; anything written between the two is completely ignored by programs that display HTML.

Spammers use HTML comments to split up a suspicious word by

inserting comments in the middle of the word. For example, *Viagra* can be broken up like this:

```
V<!--anon-->i<!--dinosaur-->
a<!--hexagon-->g<!--two-->r
<!--mouse-->a
```

That odd looking text will display as *Viagra* in any email program that understands HTML. Many spam filters will be fooled by this technique because they don't understand HTML and are unable to see the word *Viagra*. Even worse they might read the words in the comments and assume that the message is legitimate.

This is the most popular HTML trick used by spammers, and good spam filters now incorporate code that will simply strip out HTML comments before considering whether the message is spam or not. It's a simple task for a program to look for `<!--` followed by `-->` and just throw it away: after all that's what the email program is going to do when it displays the message.

In addition a spam filter can consider the very presence of HTML comments to be suspicious: after all, who sends legitimate messages like that? Spammers also sometimes use invalid HTML tags, they just make up tag names, because all browsers will just throw the tags away. Just inserting random words between `<` and `>` works as well as HTML comments:

```
V<anon>i</dinosaur>a<hexagon>g
<two>r</mouse>a
```

The Black Hole
The incredible popularity of the previous trick has been its downfall: most spam filters now strip HTML comments. But splitting up words with bits of HTML remains a spammer favourite. The *Black Hole* involves splitting up the suspicious word with spaces that have no width.

To specify the font size of a piece of text in HTML you write `<font size=X>` where `X` can be a value from
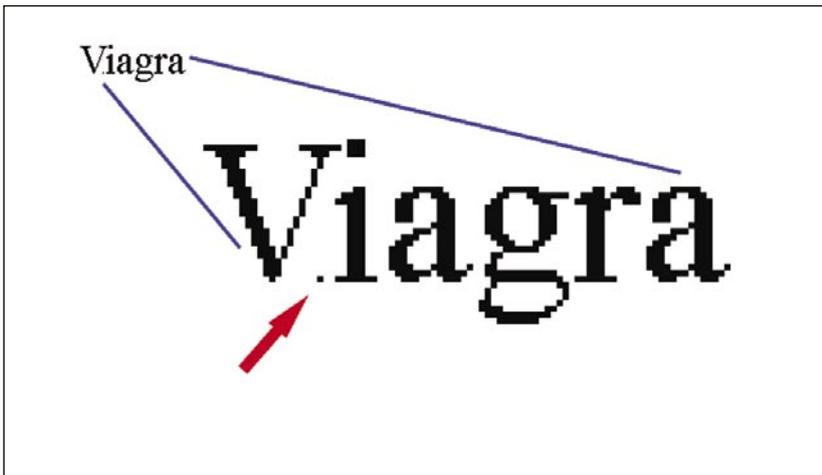
Figure 1. *Microdot*

1 to 7 (7 being the largest size and 1 the smallest). For example to say *Hello* in the smallest font available you'd write:

```
<font size=1>Hello</font>
```

Programs like Microsoft Internet Explorer and Microsoft's email programs Outlook and Outlook Express also accept the font size 0, i.e. the text has no size at all. So some spammers will put font size 0 together with a special piece of HTML syntax   which is another way of writing the space character to get a space with no width:

```
<font size=0> </font>
```

and then they use it to split *Viagra* up like this:

```
V<font size=0> </font>i
```

The arms race between spammers and anti-spammers means that up to date spam filters need to not only understand HTML comments (see the previous trick), but how HTML font sizes are specified. And once they do spammers come up with even more devious tricks: if font size 0 is going to spotted, how about font size of 1?

The Microdot
This recent innovation by spammers enables them to insert random letters in the middle of a word (thus making a spam filter that strips HTML read *Viagra* as *Vziagra* for example) and make those letters so tiny that they are almost invisible to the human eye. Welcome to the world of the microdot, or font size 1.

```
V<font size=1>z</font>iagra
```

Which when shown in an HTML-capable email program looks similar to Figure 1. As you can see the letter z has been reduced to a tiny, almost invisible dot.

Slice and Dice
The most devious form of word splitting in use involves a combination of a fixed width font and HTML tables. The crafty spammer first lays out the text using a fixed width font so that it has clearly defined columns of letters:

```
Viagra
samples
FREE
```

Then using a table with one column for each column of letters the spammer sends through the columns in order (see Listing 1 and Figure 2).

Spam filters that strip HTML tags are totally flummoxed by this technique because they end up seeing a sequence of apparently

Listing 1. *Slice and Dice*

```
<table border=0 cellpadding=0 cellspacing=0>
  <tr valign=top>
    <td><font face=Courier>V<br>s<br>F</font></td>
    <td><font face=Courier>i<br>a<br>R</font></td>
    <td><font face=Courier>a<br>m<br>E</font></td>
    <td><font face=Courier>g<br>p<br>E</font></td>
    <td><font face=Courier>r<br>l</font></td>
    <td><font face=Courier>a<br>e</font></td>
    <td><font face=Courier> <br>s</font></td>
  </tr>
</table>
```


Figure 2. *Slice and Dice*

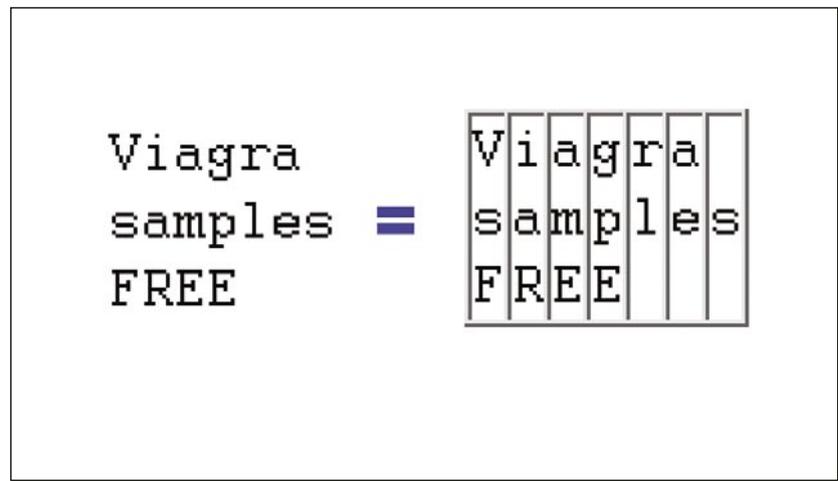**Listing 2.** *Invisible Ink*

```
<body bgcolor=white>
  Viagra
  <font color=white>
    Hi, Johnny! It was
    really nice to have
    dinner with you
    last night. See
    you soon, love Mom.
  </font>
</body>
```

Listing 3. *Camouflage*

```
<body bgcolor=#113333>
  <font color=yellow>
    Viagra
  </font>
  <font color=#123939>
    some innocent words
  </font>
</body>
```


Figure 3. *Camouflage*

random letters for analysis, as the filter is reading the text from top to bottom in columns, instead of left to right.

```
Vsf iaR ame gpe rl ae s
```

Actually reconstructing the words in the message would require the spam filter to contain some form of HTML layout engine. In practice this isn't necessary because the spammy nature of this message is given away by the use of a complex table and fixed width fonts. The actual words inside are not as important as the layout.

Inserting Hidden Good Words

Once the bad words have been hidden, spammers add words that they consider innocent. Since some spam filters have lists of good words that will allow a message through the spammer hopes that by adding some non-spam words their message will be delivered. Since the spammer doesn't want the recipient to see these good words (the recipient is meant to be concentrating on the offer for Viagra!), the spammer will arrange for the good words to be hidden from the reader but readable by spam filter.
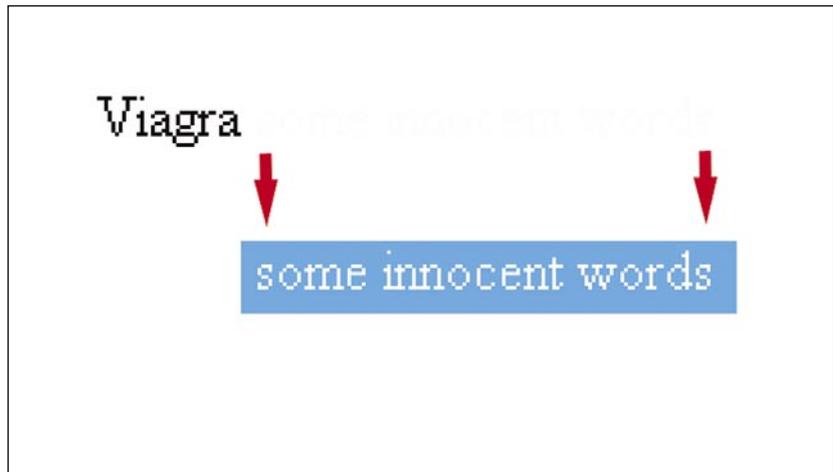
Invisible Ink
Probably the most common way of adding good words to an email is by writing white text on a white background (or any other colors where the background and foreground are the same) – see Listing 2. Spammers like this because people wont be able to read the text, but a computer that ignores color information goes ahead and reads the innocent text. Some spam filters get fooled by this, and if the spammer is clever enough to pick the right words the spam gets through.

Good spam filters understand HTML colors and spot this trick. Once spotted the email is more the likely a spam. Because of the prevalence of this trick and the fact that most good spam filters deal with it spammers have created a similar way of hiding colored text using colors that are almost, but not, identical.

Camouflage
Instead of writing white text on a white background a spammer can opt to use HTML colors that are very similar (e.g. very light grey on a white background) – see Listing 3 and Figure 3. Since colors can be specified using the hexadecimal RGB format spammers have over 16 million colors to choose from, making it possible to customize the colors individually for each spam sent.

The spammer picks a background color (in the example below #113333)

and then a foreground color that is very close to it (e.g. #123939). For the actual text that the spammer wants you to read they use a color that stands out from the background.

So the innocent words are almost invisible and the product being offered is easily read. This fools spam filters that understand the Invisible Ink trick because the colors uses are not identical, but the human eye quickly focuses on the text that is easy to read.

Good spam filters also understand this trick and calculate the similarity between the foreground and background colors using the Euclidean distance to spot text that is almost invisible to the human eye.

MIME is Money
Most email programs support the MIME encoding of messages, that allows a message to be sent in multiple parts with each part specifying its type (e.g. plain text, HTML, Word document) and will automatically display the HTML version of a message and ignore any plain text version.

Spammers exploit this by sending the innocent text in the plain portion of the message and the spam message in the HTML (see Listing 4). The spam message is displayed and the plain part is hidden from view. A spam filter that reads the entire message will end up reading both the HTML and plain versions.

Listing 4. *MIME is Money*

```
------=_NextPart_01C29D73.26716240
Content-Type: text/plain;
The modes of letting vacant farms, the duty of supplying buildings
and permanent improvements, and the form in which rent is to be
received, have all been carefully discussed in the older financial
treatises. Most of these questions belong to practical administration,
and are, moreover, not of great interest in modern times.
------=_NextPart_01C29D73.26716240
Content-Type: text/html;
<p><b>
  <font color=red>Viagra</font>
</b></p>
```

The Daily News

Another spammer favorite is to take a news story from an online site, extract the text from it and add it to their spam in the hope that the inclusion of current affairs in the message is likely to get it through.

```
<Despite statements last week from
chief U.N. inspector Hans Blix that
full cooperation was expected from
Iraq, Iraqi Foreign Minister Naji
Sabri lashed out at the United
Nations in a 19-page letter to
Secretary-General Kofi Annan
written in Arabic>
```

Obviously the spammer doesn't want you reading the news instead of their message so they wrap the text in the HTML < and > characters. Because programs that understand HTML, like email clients, ignore tags that they don't understand the text between the brackets is not displayed and ignored (the program doesn't know what the <despite> tag is in the example above).

Figure 4. *Mini Marquee*

Honorary Title

HTML provides a <title> tag that is used to set the title displayed in the web browser when the page is viewed. When the HTML is included in an email message the contents of the <title> tag are usually ignored and not displayed anywhere (the title of the email is usually its subject line).

```
<title>dinosaur reptile ghueej
egrjerijg gerrg</title>
```

So this gives the spammer yet another way to include innocent words that are never seen by the recipient.

It's Mini Marquee

The scrolling marquee tag gives the opportunity for a spammer to insert an arbitrary amount of good text and then specify that it appear in a tiny space. In the following real example the entire text scrolls by in a box 8 pixels by 8 pixels and is therefore

almost unnoticeable by the reader of the mail (see Figure 4).

```
<marquee bgcolor="white"
height="8" width="8">
  Did you ever play that game
  when you were a kid where the
  little plastic hippo tries to
  gobble up all your marbles?
</marquee>
```

Honey, I shrunk the font

And the last common trick for hiding innocent text is the use of the tiny font size 1 (see also The *Microdot* trick above). In the following example the spammer has specified font size 1 (and also the color white – so they are probably also using *Invisible Ink*) for a piece of good text but, amusingly, they seem to have failed to follow the directions in their spam software:

```
<font size="1" color="#FFFFFF">
  Random word of BIG LETTERS
  with length 1 to 22
  TSUTHRXJKVUVBECP
</font>
```

Obscuring URLs

Another way spam gets trapped by spam filters is by examining the URLs present in the message. Most spam contains at least one link to the site where the spammer is selling their wares. By building a black list of these sites its possible to delete spam based on the fact that it links to a spammer's site. Naturally spammers don't like this and try to obscure the URLs present so that a spam filter fails to match the URL against its blacklist; at the same time the spammer must ensure that the URL still works.

Enigma and Ultra

URLs are usually specified with the address of the site in a human readable form (e.g. *http://www.sophos.com*) but HTML is flexible enough to allow a variety of other forms. The following table shows four such forms: the first three all use the IP address of the
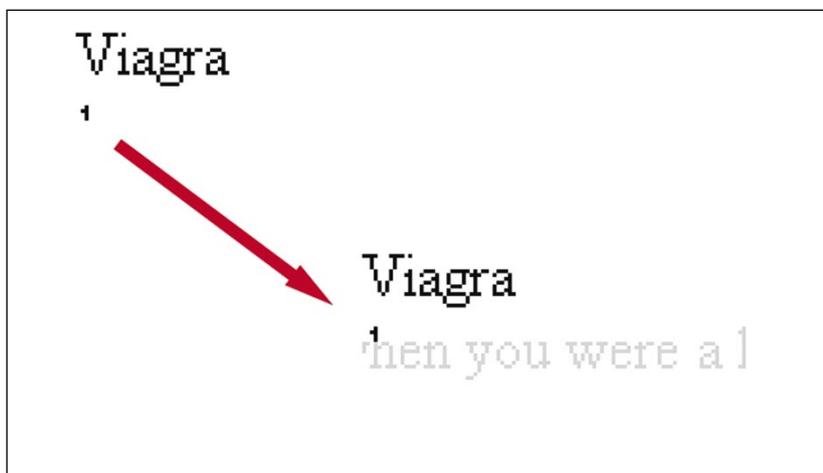
Obscuring URLs

Four methods of obscuring URLs shown in the *Enigma and Ultra* method:

- First: get the IP address of the web site (using the `host` command for example), then convert it to a single decimal number using the following formula: (X3*256^3) + (X2*256^2) + (X1*256) + X0, where IP address is X3.X2.X1.X0.
- Second: convert the number from the first method to hexadecimal, then prefix the whole by `0x`.
- Third: convert each element of the IP address to octal base, then prefix each by `0`.
- Fourth: get the ASCII value of each character in the symbolic address (using an ASCII table), convert that value to hexadecimal and prefix that value by `%`. Then concatenate into one string.

site (*www.yahoo.com*) encoded first as a single decimal number, then as a single hexadecimal and finally in dotted form using octal. The fourth form of URL uses the `%` encoding that matches each letter (or character) with its hexadecimal equivalent.

```
http://3631052355/
http://0xD86D7643/
http://0330.0155.0166.0103/
http://%77%77%77%2E%79%61%68
       %6F%6F%2E%63%6F%6D/
```

A good spam filter should understand the various codings available and reverse then before making any comparisons.

Bogus Login

A rarely used feature of URLs (at least for HTTP) is that syntax `http://username@host/` (the most common use is the simple `http://host/`). Spammers use this with randomly chosen usernames to make a suspicious URL look innocent. In this example the site being visited is not *www.microsoft.com*, but in fact the site at the IP address specified by `3631052355` (which is obscured using the trickery above for maximum effect).

```
http://www.microsoft.com@3631052355/
```

The username passed to the site is *www.microsoft.com* and will no doubt be totally ignored.

Internet Exploiter

An odd bug in Microsoft's popular Internet Explorer web browser (for which there's a patch: *http://*

www.microsoft.com/technet/treeview/ ?url=/technet/security/bulletin/MS04-004.asp) enables a crafty spammer (or scammer) to go even further than the previous two tricks by not only making the URL appear to be to a legitimate site, but also make the URL shown in the status bar of Internet Explorer appear legitimate:

```
<a href=http://www.microsoft.com
=01%01%00@3631052355>
  www.microsoft.com</a>
```

Once again this URL actually takes you to `http://3631052355/` but displays in the email program as *www.microsoft.com*, and even in the status bar shows as *www.microsoft.com* because the `%00` embedded before the `@` sign stops printing of the rest of the URL. In this trick there's an obscured spammer site, a bogus login name and explotation of a bug!

Javascript Trickery

If all those tricks weren't enough spammers also sometimes use Javascript to make their messages even more difficult to decode.

Script Writer

When using this method, the entire message to be displayed is in fact encoded inside a single variable (see Listing 5) that is not decoded until the message is opened. The spammer hopes that a spam filter wont realize the message contains a spam message and will be followed by the Javascript.

WYSI_not_WYG

Another Javascript trick is related to the obscuring of URLs. Here the real URL to be visited is hidden by changing the text that will appear in the status bar when the mouse hovers over the `ClickHere` text – see Listing 6.

The recipient is fooled into thinking they are going to one site, when in fact they are being sent somewhere totally different.

The Final Irony

The irony is that the more a spammer tries to obfuscate their message, the easier the message is to identify as a spam and delete. After all who sends real messages with trickery like *The Black Hole* or *Invisible Ink* or *The Microdot*? ∎

Listing 5. *Script Writer*

```
<HTML><HEAD><SCRIPT LANGUAGE="Javascript">
  <!-- var Words="%3CHTML%3E%0D%0A%3CHEAD%3E%0D%0A%3CTITLE%3E%3C/TITLE%3E
  %0D%0A%3CMETA%20HTTP-EQUIV%3D%22Content-Type%22%20CONTENT%3D%22text/html
  %3B%20charset%3DBig5%22%3E%0D%0A%3CMETA%20HTTP-EQUIV%3D%22Expires%22%20
  CONTENT%3D%22Sat%2C%201%20Jan%202000%2000%3A00%3A00%20GMT%22%3E%0D%0A%3C
  META%20HTTP-EQUIV%3D%22Pragma%22%20CONTENT%3D%22no-cache%22%3E%0D%0A%3C
  /HEAD%3E%0D%0A%3CFRAMESET%20ROWS%3D%22100%25%2C0%2220FRAMEBORDER%3DNO%20
  BORDER%3D%220"; function SetNewWords(){ var NewWords; NewWords =
  unescape(Words);document.write(NewWords);} SetNewWords();
```

Listing 6. *WYSI_not_WYG*

```
Remove My e-mail from my Friends Contact
<a href="http://sex.com/bPqjOL09yGCHw/"
onmouseover=  "window.status='http://%77%77%77%77.3%65%653--%69%6c11%6c%69
  --3%6c%69%6c%6l.%6f%72%67/bPqjOL09yGCHw/remove.htm';return true;"
onmouseout= "window.status=' ';return true;">ClickHere</a>
```