



The Issue

Spammers are developing more and more sophisticated methods to avoid filters. Generally, this entails attempts at sending out e-mail "waves" in which each and every e-mail is in some way unique and different from its predecessors. The relative success of each wave is then analyzed by the spammer and the resulting finds become "features" of the next spam wave.

New methods of detecting spam waves, such as extracting their core characteristics and pushing these characteristics out to clients as spam signatures, are in the final phase of development. Attempts are also being made at finding methods to predict spam changes.

Many of the filtering methods used by BitDefender have become more robust at dealing with all of the little variations encountered in spam flows. However, in 2006 there has been an increase in image spam. Simple e-mails with apparently similar images (but unique, judging by their computational differences) started polluting our inboxes in large quantities.

At the time image spam-fighting techniques were just emerging, an effective image spam detection method was that of making signatures based on the image metadata. However, given that the BitDefender antispam lab have, in the meantime, found in-the-wild spam e-mails using fresh new techniques of image poisoning intended to defeat spam filters, an entirely new technology is now needed to defeat this new development.

The Original Approach

In 2005, "image spam" accounted for approximately 10% of the total amount of spam. Such message series typically consisted of about 5-6 spam images with some minor modifications.

In recent months, however, spammers have noticed that many of the current antispam solutions are almost ineffective against this new trick so they have started attacking this niche in earnest. Image spam has increased to 30-40% of the total amount of circulating spam, with random noise changing with almost every image sent. Detection rates have dropped even further, from more than 97% to almost 65-75%.

Spam images usually contain pictures of Viagra pills, computer hardware, pornographic images, or just the classical spam message (some text and a URL) but written in a noisy image.

To do any sort of content analysis on such e-mails would mean, on the face of it, that the pictures need to be run through an optical character recognition (OCR) module. Yet common OCR filters are computationally expensive and their accuracy leaves much to be desired.

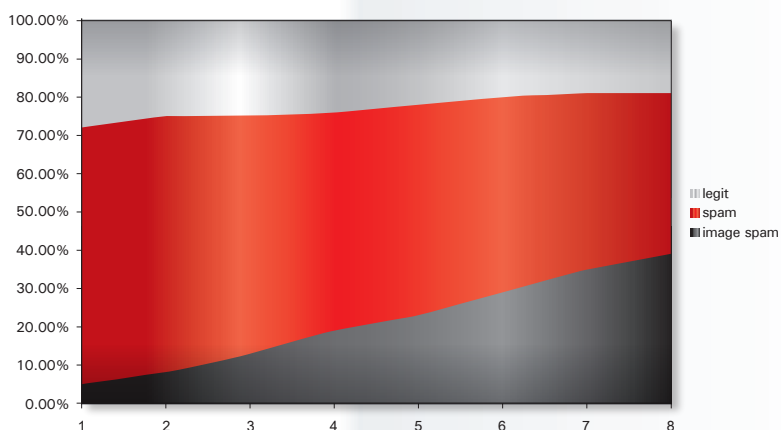


Image Spam evolution in the first 8 months of 2006



BitDefender's Approach

For a more reliable detection, BitDefender offers an alternative to OCR, namely a filter which ignores the text within images (the message, from a human point of view) and instead learns by experience some common characteristics of the images proper.

This alternative relies on the use of two techniques, histogram* extraction and histogram comparison, which have proved to be fruitful, over the time, in applications that involve image processing.

They are generally used in content-based image retrieval (e.g. extracting all pictures of dolphins from a set of vacation photos), with a rather high false positive rate. Therefore, considering them as instruments in an AntiSpam solution was quite problematic at first as false positives meant lost e-mails for the user, which was not to be taken lightly.

Experimentation has revealed that a new formula derived from these techniques, called SID (short for Spam Image Distance) can be relied upon to produce few false positives.

The Spam Image Distance algorithm picks out images based on their resemblance in point of quantity of similar colors rather than in point of shape content. From a SID perspective, for instance, although all pictures of printed pages look somewhat alike, being white or off-white, with some quantity of a darker grey, a page of the Encyclopedia Britannica does not look quite like a page of a text ad, because the proportions of white and grey are so different.

SID is used to compare images and assess the "distance" between them, which essentially means finding out how dissimilar they are.

The distances found based on the SID formula are used to compare images already included in the spam database to new images which might be spam. If the image analysis returns a score lower than a given threshold, then the image is added to the BitDefender spam images database. That is why SID is the technique of choice when dealing with spam images which are variations of other, older spam images.

While this new technique can be shown to perform well on "clean" images, there remains the problem of images having undergone obfuscation (e.g. noise adding). Fortunately, the obfuscation techniques used by spammers are well-known and the arsenal of countermeasures is similarly wide. For instance, spammers will split an image into subimages and embed them into an HTML table to reconstruct the initial image. This problem can be tackled with by stitching together the histograms of the subimages, reconstructing the histogram of the initial image and then applying a SID- based analysis on the resulting composite histogram.

Detection Rates

This patent pending technology shows a 98.7% detection rate on the BitDefender corpus of spam images (a few million samples extracted from real spam). 1.23% of these images are malformed, which means that their histograms cannot be extracted but they cannot be displayed either. A further 0.07 represent false positive results. If images that are malformed are deleted from the corpus, the detection rate quickly jumps to 100%.

With such promising results, the SID algorithm is a worthwhile addition to the arsenal of any modern antispam solution and the advances in noise reduction are expected to further improve the potential of this already very useful tool.

Common „noising“ techniques:

- Adding random pixels in the image
- Animated GIFs with noisy bogus frames
- Similar colors between different parts of the text in the image
- A long line at the end of the image (some kind of border) with random parts missing
- Splitting the image into subimages and using the table facilities in HTML to reconstruct it
- Sending different sizes of the same image
- Image poisoning - inserting legitimate pictorial content such as company logos in spam messages.
- Sending noisy legitimate pictures to confuse filters
- Sending legitimate pictures with content close to spam (e.g. mortgage images from legit mortgage companies)

*A histogram can be defined as a list of colors and their relative preponderance in an image; it indicates what colors and how many pixels of a given color exist in that image.

Contact details:

United States:
BitDefender LLC
6301 NW 5th Way, Suite 3500
Fort Lauderdale, Florida 33309

Phone: +1 (1) 8003-888062
Fax: +1 (1) 8003 888064