Combaterea imaginilor spam

În ce constă fenomenul?

Cei care lansează mesaje spam creează metode din ce în ce mai complexe de „păcălire" a filtrelor. În general, acestea se concretizează în tentative de trimitere a unor „valuri" de mesaje e-mail, în care fiecare mesaj este unic și diferit de predecesorul său. Spammer-ul analizează în ce măsură „valul" respectiv a avut succes, iar rezultatele acestei analize sunt transformate în caracteristici ale următorului „val".

În prezent, sunt pe punctul de a fi finalizate noi metode de detecție a valurilor de spam prin identificarea caracteristicilor de bază ale acestora și transmiterea lor către clienți, sub formă de semnături de spam. De asemenea, se încearcă descoperirea unor modalități de anticipare a schimbărilor survenite de la un val la altul.

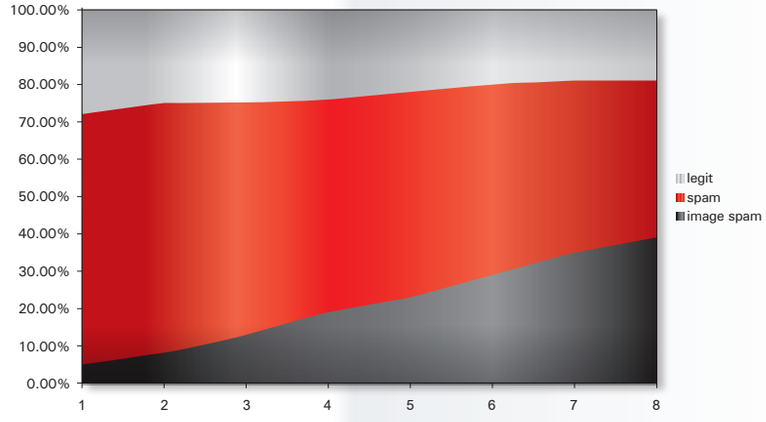
Multe dintre metodele de filtrare folosite de BitDefender și-au îmbunătățit capacitatea de identificare a variațiile subtile ale valurilor de spam. Totuși, în 2006 a existat o creștere a numărului de mesaje spam bazate pe imagini. Mesaje e-mail simple, care conțineau imagini aparent similare (dar unice datorită unor diferențe de natură computațională), au început să inunde căsuțele poștale electronice.

Atunci când tehnicile de combatere a spam-ului erau abia la început, o metodă eficientă de detectare a acestor mesaje s-a dovedit a fi cea a creării de semnături pe baza datelor suplimentare (metadata), altele decât imaginea propriu- zisă, pe care acestea le conțineau. Între timp însă, Laboratoarele antispam BitDefender au identificat mesaje spam aflate în circulație care foloseau tehnici cu totul noi de alterare a imaginilor (image poisoning) pentru eludarea filtrelor, astfel că a devenit necesară crearea unei tehnologii care să poată contracara această nouă tendință.

Abordarea inițială

În anul 2005, spam-ul bazat pe imagini, care consta, de obicei, în 5-6 imagini ușor modificate, reprezenta aproximativ 10% din totalul mesajele de acest gen.

În ultimele luni, autorii de mesaje spam și-au dat seama că multe dintre soluțiile antispam actuale au o eficiență foarte scăzută împotriva acestei noi tactici, astfel că au pornit atacuri în forță asupra acestui punct vulnerabil. Spam-ul bazat pe imagini a crescut până la 30-40% din totalul mesajelor spam aflate în circulație, fiecărei imagini trimise în acest fel fiindu-i adăugate filtre de bruiaj aleatorii. În consecință, rata de detecție a acestor mesaje a scăzut de la peste 97% la 65-75%.



Evoluția mesajelor spam bazate pe imagini în primele 8 luni ale anului 2006

Imaginile spam sunt, de cele mai multe ori, reclame pentru pastilele Viagra sau pentru componente de calculator, fotografii pornografice sau, pur și simplu, mesaje spam clasice (un text și un URL), dar scrise sub forma unei imagini bruiate.

Pentru a realiza orice tip de analiză a conținutului unor asemenea mesaje e-email, este necesară, la prima vedere, trecerea lor printr-un modul de recunoaștere a caracterelor optice (OCR). Însă, filtrele OCR obișnuite presupun un consum mare de resurse, iar acuratețea lor lasă mult de dorit.

Abordarea BitDefender

Pentru o mai mare fiabilitate a procesului de detecție, BitDefender oferă o alternativă la metoda OCR: un filtru care, în loc să analizeze textul din interiorul imaginilor (mesajul, din punctul de vedere uman), învață caracteristicile comune ale imaginilor propriu- zise.

Această alternativă se bazează pe două tehnici, obținerea și compararea histogramelor*, care au dat rezultate, până acum, pentru aplicațiile care presupun procesarea de imagini.

Aceste tehnici sunt folosite, în general, în procesul de selecție a imaginilor în funcție de conținutul lor (de exemplu atunci când trebuie să scoateți toate imaginile cu delfini dintr-o serie de fotografii de vacanță) dar au un grad de acuratețe destul de scăzut. Prin urmare, utilizarea lor ca instrumente într-o soluție Antispam a pus destul de multe probleme la început deoarece lipsa de precizie îl putea face pe utilizator să piardă mesaje e-mail legitime, un aspect deloc de ignorat.

În urma experimentelor făcute, s-a ajuns la o nouă formulă derivată din aceste tehnici, denumită SID (Spam Image Distance, adică „distanța" dintre imaginile spam), care are o rată mai mare de identificare corectă a mesajelor.

Pe baza algoritmului SID sunt alese acele imagini care se aseamănă din punctul de vedere al cantității de culori mai degrabă decât din punctul de vedere al formelor pe care le conțin. De exemplu, din perspectiva SID, deși toate imaginile de pe paginile tipărite arată mai mult sau mai puțin la fel, datorită faptului că sunt alcătuite din zone de alb și non-alb, precum și de gri mai închis, o pagină din Enciclopedia Britanică nu arată la fel cu o pagină cu un text publicitar din cauza diferenței proporționale foarte mari dintre cantitatea de alb și de gri pe care acestea le conțin.

SID este folosită la compararea imaginilor și la stabilirea „distanței" dintre ele, ceea ce înseamnă, practic, că se stabilește în ce măsură imaginile respective sunt diferite. Diferențele descoperite pe baza formulei SID sunt folosite pentru compararea imaginilor care se află deja în baza de date de spam cu imaginile suspecte, nou primite. Dacă după analizarea imaginii se obține un scor inferior unei anumite limite, atunci imaginea respectivă este adăugată la baza de date de spam a BitDefender. Acesta este motivul pentru care SID este tehnica cea mai potrivită în cazul imaginilor spam care reprezintă variații ale altora, mai vechi.

Deși această tehnică nouă se poate dovedi a fi eficientă în cazul imaginilor „curate", rămâne problema imaginilor codate (de exemplu prin adăugarea unui filtru de bruiaj). Din fericire, tehnicile de codare folosite de autorii mesajelor spam sunt bine cunoscute și arsenalul de măsuri de contracarare a acestora este pe măsură. De exemplu, pentru crearea unui mesaj spam se recurge mai întâi la divizarea unei imagini și apoi la reconstruirea ei prin introducerea fragmentelor rezultate într-un tabel HTML. Această problemă poate fi rezolvată prin alipirea histogramelor fragmentelor obținute, reconstruirea, în acest fel, a histogramei imaginii inițiale și analizarea acesteia pe baza algoritmului SID.

Rata de detecție

Această tehnologie pe cale de a fi brevetată duce la o rată de detecție de 98,7% raportată la corpusul de imagini spam al BitDefender (alcătuit din câteva milioane de mostre preluate din mesaje spam reale). 1,23% din aceste mesaje sunt malformate, ceea ce înseamnă că histogramele lor nu pot fi obținute, dar și că nu pot fi afișate. Alte 0,07% reprezintă mesaje identificate greșit. Dacă se elimină din corpus imaginile malformate, rata de detecție urcă până la 100%.

Având în vedere aceste rezultate promițătoare, algoritmul SID merită adăugat la arsenalul modern de soluții antispam. În plus, se estimează că progresele înregistrare în materie de reducere a bruiajului vor spori și mai mult potențialul acestuia.

Tehnici obișnuite de „bruiaj"

- Adăugarea de pixeli în imagine, în mod aleatoriu
- GIF-uri animate cu frame-uri false bruiate
- Culori similare în diferite porțiuni ale textului din imagine
- O linie la sfârșitul imaginii (un fel de limită) cu fragmente care lipsesc în mod aleatoriu
- Divizarea imaginii și folosirea opțiunilor de tabelare în HTML pentru reconstruirea acesteia
- Trimiterea unor versiuni de diferite mărimi ale aceleiași imagini
- Alterarea imaginii- inserarea de conținut pictural legitim, cum ar fi logo-uri de companii, în mesajele spam
- Trimiterea de imagini legitime bruiate pentru a induce în eroare filtrele
- Trimiterea de imagini legitime cu conținut asemănător cu cel al mesajelor spam (ex. imagini legate de credite ipotecare oferite de companii reale)

*)Histograma conține o listă de culori și informații despre preponderența acestora într-o imagine; aceasta indică ce culori și câți pixeli dintr-o anumită culoare există în imaginea respectivă.

Informații de Contact:

Țara: **România**
Adresă: Str.Fabrica de Glucoza, nr. 5, București
Tel: +40 21 2330780
E-mail: *sales@bitdefender.ro*
Web: *www.bitdefender.ro*

bitdefender
secure your every bit