SEODisc: Ansatz zur Erkennung von SEO-Attacken

Matthias Meyer

21. März 2011

TU Dortmund, G Data

Inhaltsverzeichnis

Was ist SEO?

- SEO = **S**earch **E**ngine **O**ptimization
- Ziel: Verbesserung der Positionierung einer Webseite in den Suchergebnissen
- Meist auf Google optimiert
- Unterscheidung in *White-Hat-SEO* und *Black-Hat-SEO*

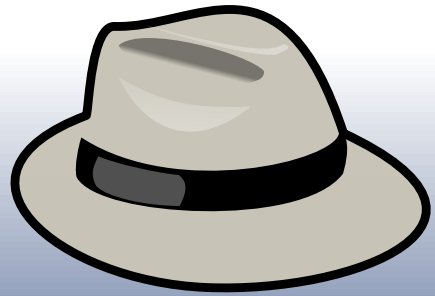
Wieso Interessant für Angreifer

- Suchanfragen Dez. 2009: [1]
  - Weltweit: ca. 131 Milliarden Suchanfragen
  - Google: ca. 87 Milliarden Suchanfragen (ca. 66 %) bzw. ca. 2.8 Milliarden Suchanfragen täglich
- Positionierung in den Top 20 Suchergebnissen
  $\Rightarrow$ viele potentielle Opfer.
- Funktionierendes Cloaking $\Rightarrow$ keine Erkennung durch Crawler
- Infektionsvektor über z.B. Drive-By-Downloads schwer durch zentrale Mittel zu filtern.
- Finanzieller Aspekt (Werbebanner) durch hohen Traffic.

[1] Quelle: ComScore - www.comscore.com - Pressemitteilung 22.01.2010

White-Hat-SEO

- On-Site
  - Meta Tags (Header, Description)
  - Suchmaschinenfreundliche URLs
  - Keywords in Überschriften verwenden
  - Vereinfachte Navigation
  - Bilder: Verwenden von ALT und Description Tags
- Off-Site
  - Link-Building



Black-Hat-SEO

- Black-Hat-SEO = White-Hat-SEO + ...
  - Content Spam
  - Keyword stuffing
  - Cloaking
    - Textfarbe
    - NoScript
    - CSS / JavaScript
    - Auswertung User-Agent
    - Auswertung Referrer
  - Link-Building
  - ...

Möglicher Angriff

1. Erstellen einer Webseite zum Thema T unter URL L_D
2. Aufbau eines SEO Netzes zum Thema T
   - Unbemerkte Übernahme existierender Webseiten $L_1, L_2, ..., L_n$ zum Thema T
   - Einfügen von Links zu allen L_i (versteckt vor menschlichen Besuchern)
   - Einfügen versteckter Links zu L_D
3. Verstecken der Manipulation durch Cloaking Techniken
   - Filtern der Seiteninhalte nach Referrer
   - Filtern der Seiteninhalte nach User Agent

Der Lösungsansatz

3 verschiedene Teilprobleme

- Finden potentiell verseuchter Webseiten (Kandidatenfindung)
- Finden von versteckten Inhalten
- Erkennen von SEO-Netzen

Verfolgter Lösungsansatz

Der Lösungsansatz

Problem

Finden potentiell verseuchter Webseiten

Annahme

- Black-Hat SEOs verwenden Schlagworte zu aktuellen Themen (z.B. Erdbeben Japan)
- Da Schlagworte aktuell vermehrt gesucht ⇒ steiler Anstieg in den Google Trends
- Verbesserung des Rankings als Folge der SEO sorgt für Position in den Top 25 Suchergebnissen.

...

Verfolgter Lösungsansatz

Der Lösungsansatz

Problem

Finden potentiell verseuchter Webseiten

Lösung

- Abfrage der Top 20 Keywords der Google Trends und Google Hot Topics (stündlich)
- Für gefundene Keywords ⇒ Googlesuche und speichere Top 25 Ergebnisse
- Gefundene URLs: Abruf der Webseite mit verschiedenen emulierten UserAgents

...

Der Lösungsansatz

Problem

Finden potentiell verseuchter Webseiten

Vorgehen

- $\sum$ 10 Abrufe einer Webseite
  - 5 Abrufe ohne Referrer Header gesetzt
  - 5 Abrufe mit Google Suche als Referrer
- Abruf jeweils mit emuliertem User Agent
  - Google Crawler
  - Firefox (3er Serie)
  - Internet Explorer (6er Serie)
  - 2x Neutraler User Agent (GTFetch)

Der Lösungsansatz

Problem

Finden von versteckten Inhalten

Annahme

- Links auf SEO-Netz Mitglieder werden durch nicht tiviale cloaking Maßnahmen versteckt
- Das Augenmerk liegt im Verstecken vor menschlichen Benutzern

Verfolgter Lösungsansatz

Der Lösungsansatz

Aufgabe

Differenzanalyse einer Webseite - Erkennen von versteckten Links

Vorgehen

- Suche Linkentsprechungen
  - $\forall$ Links $L_i \in W_{Google}$ suche Entsprechung in W_{FF}
  - $\forall$ Links $L_i \in W_{Google}$ suche Entsprechung in W_{IE}
  - $\forall$ Links $L_i \in W_{Google}$ suche Entsprechung in $W_{Neutral}$
- Suche Links in "dynamischem Content"
  - $\forall$ Links $L_i \in W_{Neutral_1}$ suche Entsprechung in $W_{Neutral_2}$
- Vergleiche Anzahl Links zwischen Referrer und None-Referrer Variante (Referrer-Analyse)

Verfolgter Lösungsansatz

Der Lösungsansatz

Aufgabe

Differenzanalyse einer Webseite -
Finden von versteckten Texten

Vorgehen

- Ermittlere relative Schlagwortdichte pro Webseitenversion
  - Berechne Schlagwortdichte von W_{Google}, W_{FF}, W_{IE}, $W_{Neutral_1}$ sowie $W_{Neutral_2}$
  - Vergleich der Schlawortdichte zwischen Versionen der Webseite
- Suche Inhalte in "dynamischem Content"
  - Vergleich Schlagwortdichte $W_{Neutral_1}$ vs. $W_{Neutral_2}$

Verfolgter Lösungsansatz

Der Lösungsansatz

Problem

Erkennen von SEO-Netzen

Annahme

- Black-Hat SEOs besitzten eigene versteckte Netzstrukturen zum pushen von infizierten Webseiten
- Diese Netzstruktur ist über versteckte Links miteinander vernetzt
- Es existieren Zyklen in dieser Netzstruktur aufgrund der möglichst vollständigen Vernetzung

...

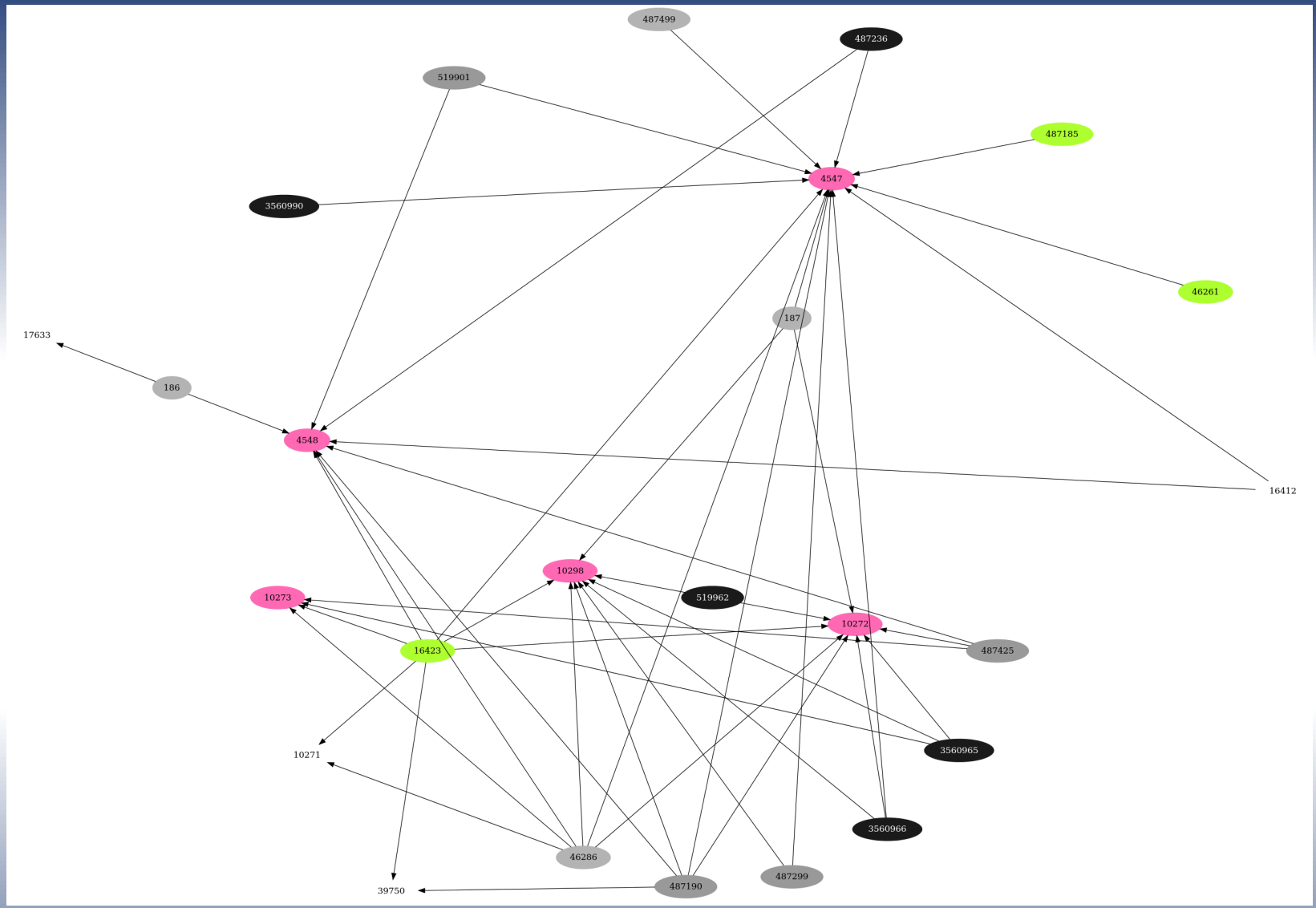
Der Lösungsansatz

Problem

Erkennen von SEO-Netzen

Lösung

- Aufbau eines Graphen mit gefundenen Links aus der Differenzanalyse
  - Knoten = URLs
  - Kanten = Links
- Suche im Graphen nach Zyklen und Zusammenhangskomponenten
- Jede Zusammenhangskomponente bildet ein SEO-Netz
- In jedem SEO-Netz suche Knoten mit hoher Zahl eingehender Kanten


487499
487236
519901
487185
4547
3560990
46261
187
17633
186
4548
16412
10298
10273
519962
10272
16423
487425
10271
3560965
3560966
46286
487299
487190
39750

Gewonnene Erkenntnisse

- Experiment durchgeführt:
  - 2 Zeiträume: 23.09.10 - 04.11.10 und 02.01.11 - 15.01.11
  - Insgesamt analysierte Webseiten: 660.749
  - 43 potentielle SEO-Netze gefunden worden
- Keine eindeutige Erkennung von schädlichen Webseiten in SEO-Netzen

Aufmerksamkeit ...

Danke für Ihre Aufmerksamkeit

Kontakt: Matthias.Meyer@tu-dortmund.de /
Matthias.Meyer@gdata.de
Infos: www.seodisc.de (comming soon)