عنوان مقاله:

A Large-scale Analysis of Content Modification by Open HTTP Proxies

هدف پژوهش:

پروکسی‌های HTTP ترافیک را بین کاربران وب و سرورهای مقصد موردنظر آن‌ها relay میکنند و به این ترتیب کاربران میتوانند به محتوایی که به هر دلیلی (محدودیت‌های ناشی از موقعیت جغرافیایی، خط مشی‌های فیلترینگ محتوا و ...) از دسترسی به آن محروم شده‌اند، دست یابند. open HTTP proxy ها امروزه بسیار مورد استقبال قرار گرفته‌اند زیرا رایگان هستند، نیازی به نصب ندارند و به راحتی قابل استفاده هستند. هدف از این پژوهش مطالعه‌ی رفتار open HTTP proxy ها با تمرکز بر شناسایی پروکسی‌های malicious است.

روش حل مسئله:

تعداد کل پروکسی‌های مورد آزمایش ۶۵۸۷۱ است که این پروکسی‌ها از وب‌سایت‌هایی که لیستی از open proxy ها را در اختیار قرار می‌دهند جمع‌آوری شدند. ده وب‌سایت crawl شدند و پروکسی‌های آن‌ها به‌صورت خودکار جمع‌آوری شدند و جمع‌آوری پروکسی از ۵ وب‌سایت دیگر به‌صورت دستی انجام شد. همچنین برای دسترسی به لیست پروکسی‌های یک وب‌سایت، اشتراک خریداری شد. پس از آن برای پی بردن به اینکه کدام پروکسی‌ها زنده[1] هستند در هر ساعت به همه‌ی پروکسی‌های مورد آزمایش بسته‌های کاوش[2] TCP ارسال می‌شود. اگر پروکسی‌ای به حداقل یک بسته‌ی کاوش پاسخ دهد یعنی زنده است پس در لیست پروکسی‌ها نگه داشته می‌شود تا مورد آزمایش قرار بگیرد. برای تست پروکسی‌ها از فریم‌ورک Selenium استفاده می‌شود و یک نمونه از مرورگر فایرفاکس launch می‌شود و به دو honeysite به‌صورت موازی درخواست فرستاده می‌شود. این دو honeysite ازنظر پیچیدگی با هم متفاوت هستند. یکی از آن‌ها یک صفحه‌ی وب کاملاً ایستا[3] است و دیگری دارای محتوای پویا[4] نیز می‌باشد و از منابع third party هم استفاده می‌کند. پس از آن‌که محتوای موردنظر fetch شد باید بررسی شود که آیا محتوا تغییر کرده است یا خیر. تشخیص تغییر محتوا برای honeysite ایستا ساده است. برای تشخیص تغییر محتوا برای honeysite پویا، آن honeysite چند بار بدون استفاده از پروکسی دانلود می‌شود و محتوای آن‌ها با هم مورد مقایسه قرار می‌گیرد و بر اساس آن یک template ایستا تولید می‌شود. سپس درخت DOM محتوای fetch شده با استفاده از پروکسی با آن قالب مقایسه می‌شود. به‌منظور دسته‌بندی رخدادهای تغییر محتوا از یک روش دوسطحی استفاده می‌شود. در ابتدا، درخت DOM محتوای fetch شده با استفاده از پروکسی، پیمایش می‌شود و موقعیت و نوع عناصر inject شده شناسایی می‌شود. سپس بر اساس تفاوت‌های بین عناصر این درخت DOM و عناصر درخت DOM مورد انتظار (template تولیدشده) دسته‌بندی در سطح اول انجام می‌شود. سپس درخت

[1] alive
[2] probe
[3] static
[4] dynamic

DOM همه‌ی نمونه‌های یک گروه با هم مقایسه می‌شوند و درصورتی‌که درخت DOM دو نمونه دقیقاً یکسان بود آن‌ها در یک دسته در سطح دوم قرار می‌گیرند.

نتایج:

برخی از وب‌سایت‌ها برای در اختیار گذاشتن لیست پروکسی‌ها نیازمند خرید اشتراک هستند و برخی دیگر این لیست را رایگان در اختیار عموم قرار می‌دهند. درصد پروکسی‌هایی که به‌درستی کار می‌کنند[5] و همچنین درصد پروکسی‌های malicious در این دودسته تقریباً یکسان است.

تنها ۳۸,۳۸ درصد پروکسی‌های زنده (پروکسی‌هایی که حداقل به یک بسته‌ی کاوش پاسخ داده‌اند) به‌درستی کار می‌کنند و می‌توانند وب‌سایت‌های مورد آزمایش را fetch کنند. اکثر این پروکسی‌ها به این دلیل کار نمی‌کنند که زمانی که می‌خواهند درخواست کاربر را forward کنند با خطاهای شبکه‌ای مواجه می‌شوند. پروکسی‌هایی که کار نمی‌کنند معمولاً برای بازه‌ی زمانی کوتاهی در لیست پروکسی‌های وب‌سایت‌ها قرار دارند و سریع حذف می‌شوند. ۳۸,۲۱ درصد از پروکسی‌هایی که به‌درستی کار می‌کنند محتوای صفحات وب بازیابی شده را تغییر می‌دهند. اکثر پروکسی‌هایی که محتوا را تغییر می‌دهند malicious نیستند و هیچ آسیبی ایجاد نمی‌کنند. به‌عنوان‌مثال بسیاری از این پروکسی‌ها tracker ها و تبلیغاتی که در صفحات بازیابی شده وجود دارند را مسدود[6] می‌کنند. تنها ۵,۱۵ درصد از این پروکسی‌ها malicious هستند که اکثر آن‌ها (۵۰,۷۹ درصد) مربوط به AS های چینی هستند. پروکسی‌های malicious اغلب کدی را تزریق می‌کنند تا به یک یا چند third party درخواست‌هایی فرستاده شود و با آن‌ها ارتباط برقرار شود. اکثر این third party ها مربوط به یکی از موتورهای جستجوی چینی هستند و حدود نیمی از پروکسی‌های malicious با این دامنه در ارتباط هستند. پروکسی‌های malicious زمان‌های خیلی کوتاه‌تری زنده هستند. همچنین سایز محتوای دانلودشده با استفاده از پروکسی‌های malicious همیشه از پروکسی‌های دیگر بیشتر است (به دلیل اینکه معمولاً یک کد جدید به آن تزریق شده است). برخی از پروکسی‌های malicious بسته به محتوا رفتار خود را تغییر می‌دهند و تنها گاهی اوقات (نه همیشه) رفتار مخرب دارند. پروکسی‌های malicious با تزریق یا تغییر محتوا اهداف زیر را دنبال می‌کنند (به ترتیب):

- تبلیغات
- جمع‌آوری اطلاعات کاربران
- دنبال کردن حرکات موس و کیبورد
- دنبال کردن با استفاده از cookie ها
- واردکردن بدافزار به ماشین کاربر

[5] properly-working
[6] block