Domain Generation Algorithms

Alessio Russo
Università degli studi di Parma

Introduzione

Un dominio è un identificatore (univoco) utilizzato per accedere a risorse su Internet. In particolare, i domini sono stringhe alfanumeriche strutturate gerarchicamente, tradotte in indirizzi IP attraverso il **Domain Name System** (DNS).

Ad esempio, il dominio *example.com* è associato a un indirizzo IP che permette ai dispositivi di localizzare il server corrispondente. I domini sono progettati per facilitare la navigazione su Internet e per essere gestibili dagli esseri umani, a differenza degli indirizzi IP numerici.

I **registrar** sono gli enti che permettono la registrazione e la gestione dei domini, mentre i **name server** sono i server responsabili della risoluzione dei nomi di dominio. Concludiamo poi introducendo i **record DNS**, delle tuple che contengono le informazioni associate a un dominio, come gli indirizzi IP (record `A` e `AAAA`), i server mail (`MX`) e i riferimenti a server di nomi (`NS`).

Classificazione dei domini

I domini possono essere classificati secondo diversi criteri.

In base alla gerarchia del DNS, si dividono in:

- **top-level domains** (TLD) come `.com`, `.net`, `.org`, `.gov` e `.edu`.
- **second-level domains** (SLD) come `example` in `example.com`.
- **subdomains** (o *third-level domains*) utili per l'organizzazione delle risorse, come `blog` in `blog.example.com`.

Sulla base dell'accessibilità, si dicono invece:

- **generici** (gTLD), registrabili pubblicamente, come `.com`, `.net` e `.info`.
- **nazionali** (ccTLD), assegnati a specifiche nazioni, come `.it` per l'Italia o `.fr` per la Francia.
- **sponsorizzati** (sTLD), gestiti da specifiche organizzazioni (o settori), come `.gov` per enti governativi e `.edu` per istituzioni educative.
- **riservati**, non registrabili pubblicamente, come `.localhost` e `.test`.

Infine, in base allo scopo per cui vengono registrati, si distinguono in:

- **legittimi**, registrati e utilizzati per siti web, aziende e servizi autentici.
- **malevoli**, generati da malware o cybercriminali per attacchi informatici.
- **parcheggiati**, registrati ma non utilizzati attivamente, spesso con l'intento di rivenderli o per protezione del marchio.

Algoritmi di generazione

Gli algoritmi di generazione di dominio, noti come **DGA** (Domain Generation Algorithms), sono strumenti automatici per la creazione di domini, sulla base di schemi predefiniti.

Originariamente sviluppati per l'utilizzo in applicazioni legittime, come la creazione di indirizzi temporanei o la gestione di servizi distribuiti, questi algoritmi vegono oggi largamente utilizzati dal cybercrimine per eludere i sistemi di sicurezza e mantenere attive le comunicazioni tra malware e i server di comando e controllo (C2).

Il loro funzionamento segue, generalmente, lo schema qui riportato: un malware infetto esegue l'algoritmo che, a intervalli regolari, genera una lista di possibili domini, a cui tenterà di connettersi per ricevere istruzioni. Il server C2, a sua volta, registrerà anticipatamente uno dei domini generati, permettendo così la comunicazione tra l'attaccante e il sistema compromesso.

Questi algoritmi sfruttano funzioni deterministiche o pseudo-casuali per generare (in maniera prevedibile) un insieme di domini che cambiano nel tempo, a partire da uno specifico *seed* (come una data, una chiave segreta o un valore ottenuto da una risorsa pubblica).Quando un dominio viene bloccato dalle misure di sicurezza della rete, l'algoritmo continua a generare nuovi indirizzi, rendendo il rilevamento e la mitigazione estremamente complesse.

Topologie di algoritmi

Gli algoritmi di generazione dei domini possono essere classificati in diverse categorie, a seconda della logica impiegata per la creazione dei nomi.

Gli algoritmi **basati su dizionario** combinano parole predefinite per generare nomi di dominio che risultano piuttosto realistici, e perciò difficili da distinguere da quelli legittimi. Questi algoritmi offrono quindi il vantaggio di generare domini apparentemente naturali, ma la loro efficacia dipende dalla dimensione e dalla varietà del dizionario utilizzato.

Un'altra tipologia comune di DGA è rappresentata dagli algoritmi **pseudo-casuali**, che generano domini(spesso stringhe alfanumeriche casuali), basandosi su un *seed* (generalmente temporale) per ottenere risultati riproducibili. Sebbene questi domini possano variare dinamicamente, la loro natura casuale li rende più facili da rilevare con strumenti di analisi automatica.

Alcuni algoritmi, più sofisticati, adottano **modelli statistici** per creare domini che imitano quelli reali. Questi metodi utilizzano spesso modelli Markoviani o altre tecniche di apprendimento statistico per produrre nomi che sembrano naturali, aumentando così la difficoltà di rilevamento.

Gli approcci più avanzati si basano invece su tecniche di **apprendimento automatico** e reti neurali per generare domini che si adattano dinamicamente ai protocolli di sicurezza, risultando quindi in sistemi sempre più difficili da individuare. Sebbene questa soluzione offra un elevato livello di resistenza alle misure di sicurezza, la sua implementazione spesso complessa, oltre a richiedere maggiori risorse computazionali.

DGA nelle botnet e nei malware

Le botnet sono reti di dispositivi compromessi, controllati in modo remoto tramite un'infrastruttura C2. Un elemento fondamentale per il "corretto" successo di una botnet è la capacità di comunicare con il server C2 senza essere intercettata o bloccata. Per questo motivo, i malware moderni utilizzano DGA per generare dinamicamente nomi di dominio difficili da prevedere. Questa strategia ha un duplice vantaggio: da un lato, permette agli attaccanti di registrare solo alcuni dei domini generati, riducendo i costi e la visibilità; dall'altro, rende più difficile per le forze dell'ordine e i ricercatori di sicurezza bloccare l'infrastruttura C2.

Conficker

Uno dei primi (e più noti) esempi di malware che ha impiegato un algoritmo di generazione dei domini è conficker. Scoperto nel 2008, il suo DGA generava ogni giorno circa 250 domini diversi, basandosi su un seed condiviso

tra tutte le istanze del malware. Questo approccio permetteva al botmaster di registrare, secondo necessità, uno di quei domini, e fornire istruzioni ai dispositivi infetti. Successive varianti di Conficker hanno ampliato il numero di domini generati fino a 50.000 al giorno, rendendo più difficile per i sistemi di sicurezza la loro identificazione e il loro blocco.

Il principale limite di Conficker era che il suo algoritmo di generazione era relativamente semplice, e una volta identificato il metodo di creazione, era possibile prevedere i domini futuri e bloccarli preventivamente.

Gameover Zeus

Gameover Zeus è stato un altro malware significativo che ha utilizzato DGA per eludere i sistemi di sicurezza. A differenza di Conficker, che generava i domini in modo relativamente semplice, Gameover Zeus impiegava un DGA più sofisticato che combinava fonti di entropia esterne, come i titoli di importanti testate giornalistiche, per generare domini difficili da prevedere. Questo tipo di tecniche rende complessa la creazione di blacklist efficaci, poiché i domini variavano in base a eventi non controllabili.

Necurs

Necurs è una botnet che ha utilizzato un DGA avanzato per evitare il rilevamento. Questo malware ha introdotto un metodo di generazione basato su un modello statistico, producendo nomi di dominio apparentemente naturali, e quindi più difficili da individuare con filtri euristici (e.g. basati sull'entropia del nome di dominio).

Questa botnet è stata una delle più longeve e utilizzate per il malware finanziario, facilitando la diffusione di ransomware come Locky e banking trojan come Dridex.

Matsnu e suppobox

Questi due malware sono esempi di DGA basati su dizionario. A differenza degli algoritmi pseudo-casuali che generano stringhe alfanumeriche casuali, Matsnu e Suppobox utilizzano parole prese da un dizionario per produrre domini apparentemente autentici. Questo approccio riduce la probabilità che il traffico verso questi domini venga classificato come sospetto, poiché i nomi generati appaiono simili a quelli utilizzati per siti legittimi.

Uno degli elementi chiave che ha reso questi algoritmi così efficaci nel campo del cybercrimine è la loro capacità di adattarsi ai cambiamenti nei sistemi di difesa. Quando una tecnica viene identificata e bloccata, gli sviluppatori di malware affinano i loro DGA per renderli più sofisticati e resistenti alle contromisure.

Tecniche di rilevamento

Il rilevamento dei domini generati tramite DGA prevede diversi approcci

Analisi lessicale e sintattica

Uno dei primi metodi utilizzati per identificare i domini generati algoritmicamente è l'analisi delle loro caratteristiche **lessicali** e **sintattiche**. Come anticipato, infatti, molti DGA, generano domini privi di significato, spesso caratterizzati da lunghe stringhe alfanumeriche casuali. Gli algoritmi di rilevamento basati su questa tecnica esaminano quindi parametri come la lunghezza del dominio, la frequenza delle lettere, la presenza di sillabe comuni, l'entropia della stringa e la distribuzione dei caratteri.

Ad esempio, un dominio generato da un DGA pseudo-casuale come `axwscwsslmiagfah.com` appare molto diverso da un dominio legittimo come `unipr.it`. Utilizzando metriche statistiche, è spesso possibile classificare un dominio come sospetto.

Analisi del traffico di rete

Un altro approccio efficace per il rilevamento dei DGA è l'analisi del comportamento del traffico di rete. A differenza dei domini legittimi, che vengono interrogati da un ampio numero di utenti, i domini generati dai malware presentano spesso comportamenti anomali.

Un primo indicatore è l'elevata velocità di risoluzione dei DNS. Un dispositivo infetto da malware che utilizza un DGA tenterà di connettersi a centinaia o migliaia di domini in pochi secondi, nella speranza di trovare un dominio C2 valido. Questo comportamento è diverso da quello di un utente normale, che visita un numero relativamente basso di domini in un determinato periodo di tempo.

Inoltre, poiché i domini malevoli tendono a essere attivi per periodi brevi, il monitoraggio della **longevità dei domini** può quindi diventare un valido indicatore per le attività malevole

Approcci di machine learning

Con l'aumento della complessità dei DGA, le tecniche tradizionali di rilevamento, basate su regole euristiche, si sono dimostrate sempre meno efficaci. Per questo motivo, i sistemi di sicurezza più complessi implementano modelli basati sull'apprendimento automatico per individuare pattern ricorrenti nei domini generati.

L'apprendimento automatico permette infatti di addestrare modelli su grandi quantità di dati, distinguendo tra domini legittimi e domini generati algoritmicamente. I modelli di classificazione più utilizzati sono basati su reti neurali, alberi decisionali e support vector machines.

Un esempio di tecnica basata sul machine learning prevede l'uso di reti neurali convoluzionali per analizzare la struttura dei nomi di dominio e identificare somiglianze con campioni noti di DGA. Inoltre, l'utilizzo di algoritmi di clustering permette di raggruppare i domini con comportamenti simili, evidenziando quelli che potrebbero essere generati automaticamente.

Contromisure

Oltre al rilevamento dei DGA, esistono diverse strategie di prevenzione che possono ridurre l'impatto di queste minacce e aiutano a proteggere le reti.

Sinkholing

Una delle misure più efficaci è il sinkholing, che consiste nel registrare preventivamente i domini generati dagli algoritmi noti e reindirizzare il traffico malevolo verso server controllati, anziché verso il server C2 del malware. In questo modo, è possibile impededire al malware di ricevere nuovi comandi.

Questa tecnica è stata ampiamente utilizzata contro Conficker

Blacklist dinamica

Un'altra strategia spesso utilizza è la realizzazione di una blacklist dinamica, che aggiorna periodicamente l'elenco dei domini sospetti, identificati tramite i metodi di rilevamento sopra descritti. Tuttavia, poiché i malware generano continuamente nuovi domini, questo approccio richiede aggiornamenti costanti per rimanere efficace.

Filtraggio DNS

Un metodo più avanzato è il DNS filtering, basato su intelligenza artificiale, che combina tecniche di machine learning e analisi del traffico per identificare in tempo reale i domini malevoli.

Questi sistemi possono essere implementati a livello di firewall , o direttamente all'interno di

sistemi DNS, per bloccare automaticamente le richieste verso domini sospetti.

Difesa proattiva

Infine, è fondamentale adottare strategie di difesa proattiva, come:

- il **monitoraggio continuo** del traffico DNS, che permette di identificare anomalie nel traffico e rispondere tempestivamente a eventuali attacchi.
- la **segmentazione della rete**, che impedisce ai malware di propagarsi all'interno della rete.

Implementazione

Di seguito è riportato un esempio, in Python, di DGA basato su una funzione pseudo-casuale. L'algoritmo prende come input una data (anno, mese e giorno) e restituisce un nome di dominio generato dinamicamente.

L'idea alla base di questo codice è quella di creare una sequenza pseudo-casuale di caratteri, in base ai valori della data fornita.

```
def generate_domain(year: int, month:
    int, day: int) -> str:

"""Generate a domain name for the given
    date."""
domain = ""

for i in range(16):
    year = ((year ^ 8 * year) >> 11) ^
        ((year & 0xFFFFFFF0) << 17)
    month = ((month ^ 4 * month) >> 25)
        ^ 16 * (month & 0xFFFFFFF8)
    day = ((day ^ (day << 13)) >> 19) ^
        ((day & 0xFFFFFFFE) << 12)
    domain += chr(((year ^ month ^ day)
        % 25) + 97)

return domain + ".com"
```

In particolare, ogni iterazione del ciclo modifica i valori di `year`, `month` e `day` utilizzando operazioni bitwise (`shift`, `XOR` e `AND`) per manipolare i valori numerici e trasformarli in caratteri alfabetici.

Poiché il ciclo viene eseguito 16 volte, il risultato finale sarà una stringa di 16 caratteri alfabetici, seguiti dal suffisso `.com`.

Ad esempio, l'esecuzione del codice, dando come input il 7 gennaio 2014, produce in output il dominio

`intgmxdeadnxuyla.com`

Conclusioni

Il contrasto ai DGA è una sfida continua per la cybersecurity, poiché gli attaccanti migliorano costantemente i loro algoritmi per sfuggire ai sistemi di rilevamento. Tuttavia, l'evoluzione delle tecniche di analisi del traffico, l'uso del machine learning e l'adozione di strategie di mitigazione permettono di ridurre significativamente l'efficacia di questi attacchi. Per garantire una protezione efficace, è essenziale combinare più metodi di rilevamento e difesa, aggiornando costantemente le strategie di sicurezza per affrontare le nuove generazioni di DGA.