IL PROBLEMA DELL'IDENTIFICAZIONE NELLA *NETWORK FORENSICS*

di Marco Mattiucci e Raffaele Olivieri

La *Network Forensics* (NF), nell'ambito del più generale *Digital Forensics*, si occupa della cattura, registrazione ed analisi di comunicazioni di rete (impostate su precisi protocolli standard) al fine di ottenere informazioni utili allo svolgimento di indagini tecniche in vari ambiti legali[1] (aziendali interni, penali, civili, ecc.). Data la diffusione del protocollo IP (*Internet Protocol*) ad oggi la NF tende ad identificarsi con la cattura, registrazione ed analisi di pacchetti IP ma non si limita a questo. **Uno degli scopi fondamentali e particolarmente ricorrenti di tali analisi è il tracciamento dei pacchetti e l'identificazione a vari livelli: posizione geografica, autore, responsabilità, e svariati altri aspetti**. Il grande volume di attività degli utenti su *web*, generalmente effettuata tramite *browser*, scatena tutta una serie di comunicazioni IP, anche molto sofisticate, che hanno spinto lo sviluppo di particolari studi di NF (non limitati al *web*) codificati in una importante sotto-materia denominata *Internet Forensics*, di anch'essa si parlerà in questo lavoro.

Prima di procedere è bene spendere alcune parole su Internet, la *Interconnected Network of Networks* di portata mondiale, sempre limitatamente agli scopi (forensi) di questo intervento. Internet si può ben considerare come un insieme di servizi distribuiti costruiti sul paradigma *client-server* che fondano sulla comunicazione IP, supportata dall'architettura della Rete delle reti, il tramite principale delle "transazioni" di richiesta e soddisfazione dei compiti. L'accesso dell'utente comune al *web* è appunto un esempio di attività di queste applicazioni distribuite: il *client* (*browser* di navigazione), preposto all'invio delle richieste di servizio, ed il (o i) *web server*, che usualmente è in esecuzione su un altro sistema (o insieme di sistemi) comunque connesso ed è preposto appunto alla fornitura del servizio. Uno *stack* di protocolli viene coinvolto in tutto questo ed una classica modellizzazione delle interazioni sistemiche è quella per strati come descritto nel modello a 5 livelli[2] tipico delle reti *IP based*.

L'analisi che si rende possibile sugli strati visti, ai fini legali è classificabile a seconda di tempo e sorgente dei dati, in particolare:

- ***a.*** ***Real time***: cattura dei pacchetti e/o dei *frame* con valutazione istantanea dei contenuti (*net monitoring*);
- ***b.*** ***Post mortem***: cattura dei pacchetti e/o frame e/o messaggi a livello applicazione con valutazione dopo il fatto;
- ***c.*** ***Flow based***: analisi delle meccaniche di flusso (statistiche, log, ecc.);
- ***d.*** ***Content based***: cattura selettiva dei pacchetti in base a parametri di contenuto.

La (a) è un'attività tipica del *network security* (es. degli IDS - *Intrusion Detection System*[3]) e delle intercettazioni telematiche mentre la (b) è la tipica applicazione del NF. La (c) è un'indagine di massima su grandi sistemi di comunicazione mentre la (d) è tipica delle intercettazioni parametriche. **In tutte queste attività la capacità di poter associare gli elementi di comunicazione a degli utenti, delle posizioni fisiche, dei numeri di telefono, dei *MAC address* e più in generale a precisi riferimenti nel mondo reale è il cosiddetto problema dell'identificazione**.

La menzionata stratificazione ha snellito notevolmente sia la fase di progettazione sia quella di sviluppo delle reti ma ha introdotto degli ineluttabili elementi di debolezza dal punto di vista dell'identificazione, poiché ogni strato è costretto a fruire di servizi in "*black box*" ignorando cioè totalmente le modalità di fornitura del servizio (principio di incapsulamento degli strati).

Un esempio classico è la possibilità dell'attacco "*Man-In-The-Middle*" consistente, per un osservatore, nell'infraporsi tra due *host* ed intercettare, leggere, inserire o modificare a piacere, e replicare messaggi tra le parti, senza che questi avvertano la sua esistenza ed intuiscano la reale provenienza ed integrità dei messaggi. Un altro caso ricorrente è quello del "*DNS hijacking*" o "*DNS redirection*" in cui vengono alterati i risultati del DNS, ad opera, ad esempio di un *malware*, ridirezionando le richieste di risoluzione dei nomi.

È opportuno fare delle ipotesi quando si opera cercando di affrontare il problema dell'identificazione sulle reti:

1. lo stesso utente può fare arrivare alla macchina *target* più richieste da IP differenti in maniera coordinata;
2. i *router* hanno limitata capacità di elaborazione e di immagazzinamento dei pacchetti;
3. i *router* raramente sono compromessi;
4. non tutti i *router* possono fornire informazioni utili all'identificazione;
5. minore è il numero di pacchetti che formano una comunicazione minore è la possibilità di identificazione.

Senza queste ipotesi è possibile lo stesso svolgere delle indagini ma il livello di complessità è tale che a giustificarne lo sforzo in termini di risorse devono esserci fondi economici notevoli e motivazioni oltre la media.

Si può passare ora ad un caso pratico che rispetta le suddette ipotesi per vedere come si opera. Trattando di *Web Forensics* si può seguire il percorso che subisce la tipica richiesta di una pagina *web* mediante un *URL* scritto da un utente su un generico *browser* di navigazione[4]. Il *browser* invierà un messaggio GET (*Hypertext Transfer Protocol* - HTTP) di richiesta al *web server* associato all'URL da noi richiesto. Questo, a livello applicazione, è possibile perché tale richiesta è codificata da opportune librerie del sistema operativo. L'URL, che è una stringa di caratteri alfanumerici, deve essere interpretata al fine di ottenere un *IP address* ed un numero di porta del destinatario (il *web server*). In questo processo è il *Domain Name System* (DNS) che si occupa della "traduzione". Successivamente entreranno in ballo altre traduzioni ancora, ad esempio correlate al *MAC*

address (indirizzo della scheda di rete). Pertanto, in maniera sintetica e semplificata:

- viene generato un messaggio per interrogare il DNS al fine di "risolvere" l'URL richiesto;
- il messaggio viene inviato al livello "trasporto", incapsulato in segmenti numerati ed associato un *socket*;
- il messaggio passa al livello "rete", vengono creati i pacchetti con l'indirizzo IP del DNS da raggiungere, (spesso già presente nella configurazione di rete dell'elaboratore in uso);
- a livello "collegamento" vengono creati i *frame* incorporando i *MAC address* della scheda di rete di origine (la macchina stessa) e quella di immediata destinazione (es. un *router* ADSL);
- a livello "fisico" partono i segnali elettrici equivalenti ai *frame* appena strutturati;
- dopo una serie di salti tra *MAC address* diversi viene raggiunta la scheda di rete del DNS che opera all'inverso di quanto visto, inserisce la richiesta in coda di attesa, appena possibile risolve il quesito (non è detto che sia un solo *server* a svolgere questa complessa funzione) per poi spedire al mittente la risposta con passi molto simili a quanto già visto.

Nell'ipotesi che tutto sia andato a buon fine, può iniziare il processo di comunicazione con il *web server*. Si deve sottolineare che, mentre la comunicazione con il DNS avviene in UDP/IP, quella con il *web server* è in TCP/IP e quindi è più articolata:

- il *browser* attiva una connessione TCP con l'*host* avente *IP address* appena ricevuto, mediante una complessa fase di accordo ("*handshake*");
- sfruttando questa connessione viene richiesta la pagina (corrispondente all'URL) al *web server*;
- la pagina *web* viene inviata al *browser* (*client*);
- il *browser* dopo aver svolto un'analisi (*parsing*) della pagina *web* richiede tutte le risorse in essa contenute.
- per ogni risorsa individuata possono scatenarsi ulteriori richieste che potrebbero ovviamente conivolgere anche e nuovamente il DNS ed altri *web server*.

Verificare dal punto di vista forense, anche solo per il problema dell'identificazione, i dati che sono circolati su un canale relativamente ad una navigazione *web* risulta quindi arduo perché ognuno degli indirizzi considerati (IP del/dei DNS, IP del/dei *web server*, porte, MAC, …) ad ognuno dei livelli coinvolti può essere soggetto ad alterazione e ad ogni modo, concorre alla soluzione del problema di identificazione.

Questo vuole dire che i dati eventualmente prelevati dal canale (*sniffing*) devono essere analizzati a diversi livelli di astrazione (*frame*, pacchetti, segmenti) e in ognuno di tali livelli devono essere estratti e valutate le intestazioni (*header*) al fine di verificare la coerenza dell'indirizzamento.

L'esistenza fortunatamente di NFAT (*Network Forensic Analysis Tool*) disponibili in varie modalità, anche *online* e commercialmente, permette di affrontare tale problema compiutamente anche se bisogna dire che **l'indagine in se è ancora poco automatizzabile ed i criteri di ricerca della coerenza di indirizzi e contenuti sono di natura empirica e basati sulla capacità di osservazione dell'analista.**

E' bene tenere conto del fatto che a livello forense non vengono studiate solo comunicazioni *web* ma diverse altre tipologie di protocolli e questo aumenta ancora la complessità dello studio (si pensi al problema di identificazione di un chiamante di Skype o di un utente di Facebook). Ad ulteriore incremento della problematica vi è la necessità di studiare dei veri e propri attacchi su sistemi in rete svolti mediante complesse articolazioni di TCP/IP e UDP/IP[5] (non è strano oggi parlare di protocolli di attacco), ad esempio in un attacco distribuito lo stesso *target* può essere raggiunto da migliaia di pacchetti provenienti da *IP address* eterogenei e corrispondenti a posizioni fisiche molto distanti tra loro nonché apparentemente scorrelate. In quest'ultimo caso è necessario implementare complessi protocolli di protezione dei dati come il *packet-marking* e/o il *packet-logging* al fine di poter essere efficaci a livello forense ad attacco avvenuto e poter effettuare così un valido tracciamento (*IP traceback*).

Per quanto riguarda, comunque, l'*Internet Forensics* più comune, il problema dell'identificazione si concentra sullo studio soprattutto degli *IP address*. La catena di identificazione[6] in tal senso passa attraverso gli ISP (*Internet Service Provider*), i CSP[7] (*Cloud Service Provider*) e le diverse autorità preposte, a livello internazionale, alla distribuzione degli *IP address* (*Regional Internet Registry* - RIR). **I settori di applicazione di tali indagini sono generalmente: tracciamento *email*, P2P, siti *web*, IM, *social networking*[8]**. Le problematiche di identificazione possono essere complesse anche in questi settori ormai definiti classici in quanto alla presenza su Internet di: *proxy anonimyzer, remailer, rewebber*, limitata affidabilità dei *DB whois*, indisponibilità di *file* di *log* e non cooperazione da parte di società che vendono come servizio proprio l'anonimato.

Il problema di identificazione ha poi diversi risvolti anche di natura legale e non solo tecnica, primo fra tutti la possibilità di associare un preciso utente, e non solo una macchina o un punto di collegamento, ad un'azione o comunicazione avvenuta su una rete. Tale associazione può avvenire a seguito di particolari evidenze rinvenute sulla macchina del presunto autore (attività di *computer forensics*) o a seguito dell'impiego, da parte dell'utente, di specifiche *password* ma, al momento spingersi oltre queste osservazioni significa eesporsi a valide contestazioni della controparte. ©

RIFERIMENTI BIBLIOGRAFICI

1. Sul Network Forensics http://www.marcomattiucci.it/networkforensicsarea.php;
2. Internet e Reti di Calcolatori - di James F. Kurose e Keith W. Ross (Brossura - 1 feb. 2003);
3. Sugli Intrusion Detection System http://www.marcomattiucci.it/ids.php;
4. Digital Forensics Magazine - Issue 1 (2009/2010) - "Anatomy of a web-request" (T.Watson);
5. Digital Forensics and Cyber Crime - Proceedings of the 2nd international ICST Conference ICDF2C 2010 - "An IP Traceback Model for Network Forensics" (E.S.Pilli, R.C.Joshi, R.Niyogi);
6. Rassegna dell'Arma dei carabinieri n. 3 - Anno 2009 - "Internet Forensics" (M.Mattiucci, G.Delfinis);
7. Advances in Digital Forensics VII - Proceedings of the VII IFIP WG international conference on Digital Forensics 2011 - "Cloud Forensics" (K.Ruan, J.Carthy, T.Kechadi, M. Crosbie);
8. Digital Forensics and Cyber Crime - Proceedings of the 1st international ICST Conference ICDF2C 2009 - "Data Mining IM Communications to perform author identification for cybercrime investigation" (A.Orebaugh, J.Allnutt).◊