Co skrývá síť BitTorrent?

Technická zpráva FIT VUT v Brně

Libor Polčák

Technická zpráva č. FIT-TR-2018-05
Fakulta informačních technologií, Vysoké učení technické v Brně

Naposledy změněno: 13. prosince 2018

Co skrývá síť BitTorrent?

Libor Polčák

Vysoké učení technické v Brně, email: polcak@fit.vutbr.cz

Abstrakt Síť BitTorrent je v současné době nejčastěji používanou overlay sítí pro výměnu souborů. Z pohledu bezpečnostního výzkumu se nabízí několik výzkumných otázek. Tato technická zpráva se zabývá jejich hledáním. Pro snadnější orientaci a seznámení čtenáře s problémem obsahuje zpráva také základní terminologii související se síti BitTorrent.

1 Úvod

Síť BitTorrent patří mezi peer-to-peer sítě pro výměnu souborů. Síť BitTorrent je decentralizovaná, obsah v ní může nabízet kdokoliv. Síť BitTorrent je využívaná pro distribuci obsahu v souladu s autorským právem jako např. distribucemi GNU/Linux[1]. Na druhou stranu se síť BitTorrent používá i pro sdílení obsahu, ke kterému nemají nabízející autorská práva[2].

Cílem tohoto dokumentu je popsat jednotlivé případy použití sítě BitTorrent zaměřené na monitorování této sítě v souvislosti s bezpečnostním výzkumem. Technická zpráva také poskytuje reference na ostatní dokumenty vzniklé v souvislosti s projektem *Integrovaná platforma pro zpracování digitálních dat z bezpečnostních incidentů* v roce 2018 a s výhledem na rok 2019.

Sekce 2 poskytuje potřebnou terminologii potřebnou ve zbytku technické zprávy. Sekce 3 vyjmenovává jednotlivé případy použití sítě BitTorrent se zaměřením na bezpečnostní výzkum, přínosy projektu *Integrovaná platforma pro zpracování digitálních dat z bezpečnostních incidentů* v roce 2018 a plány pro rok 2019. Práci shrnuje sekce 4.

2 Terminologie protokolu BitTorrent

Tato sekce se zabývá potřebnou terminologií nutnou k porozumění sekcím následujícím.

Data nabízená v síti BitTorrent mohou obsahovat buď jeden konkrétní soubor, nebo celou adresářovou strukturu obsahující libovolné množství souborů. Nabízející rozdělí soubor do malých dílů, typicky o velikosti 256 kB [3] — *piece*, které mohou být následně mezi jednotlivými peery vyměňovány samostatně. V okamžiku, kdy některá ze stanic získá obsah jednoho dílu, začne jej nabízet

[1] https://thepiratebay.org/rss/top100/303, https://torrent.fedoraproject.org/

[2] Např. https://thepiratebay.org/rss/top100/201

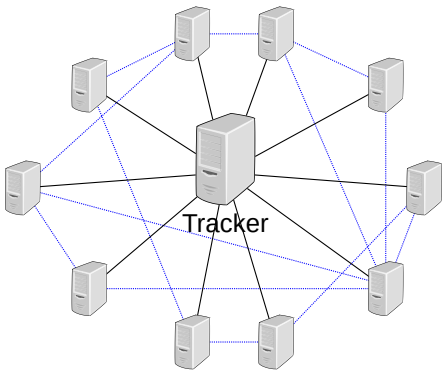
ostatním zájemcům. Každý ze stahujících tedy současně nabízí již staženou část obsahu ostatním stahujícím.

Všechna nabízená data v síti BitTorrent je možné popsat *informačním slovníkem* (info dictionary), který mimo jiné obsahuje následující položky [3]:

piece length udává velikost jednoho dílku.
pieces je posloupnost hašů SHA-1 pro ověření správnosti stahovaných dat.
path je doporučená cesta, do které mají být data stažená.

Protože je tento informační slovník použit jako vstup pro výpočet tzv. infohaše (info_hash), který slouží jako základní identifikátor stahovaných dat, může prvotní nabízející ovlivnit podobu infohaše. Neexistuje tedy jednoznačná převodní funkce mezi konkrétním obsahem a infohašem. Prvotní nabízející může tedy výsledný infohaš ovlivnit například změnou doporučeného názvu souboru a počtem dílů, na které se bude výměna dat dělit.

Původní využití protokolu BitTorrent předpokládalo existenci specializovaných serverů, tzv. *trackerů*. Každý aktivní uzel musí pro všechna nabízená i stahovaná data znát adresu trackeru. Tracker si pamatuje pro obhospodařované infohaše aktivní uzly a hraje tak roli prostředníka — libovolný uzel se jej může zeptat na identitu ostatních peerů mající zájem, či nabízející data identifikovaná poptávaným infohašem. Obrázek 1 ukazuje tzv. *swarm* — množinu peeru sdílejících mezi sebou jeden obsah identifikovaných infohašem. Aby se mohl nový zájemce do swarmu přidat, musí nejdříve kontaktovat tracker a od něj se dozvědět o existenci ostatních peerů.



Obrázek 1. Tracker hraje roli ústředního uzlu v připojování nových peerů do swarmu (modré, přerušované čáry).

Aby mohl uživatel využívající síť BitTorrent používat, musí tedy ve variantě s využitím trackerů znát jak infohaš, tak kontaktní informace jednoho, či více

trackerů. Pro šíření těchto informací slouží tzv. *metasoubory* (meta files), častěji známé jako soubory *.torrent* [3]. V rámci souborů .torrent je uložen informační slovník a další parametry včetně infohaše a trackeru. Soubory .torrent používají speciální kódování, tzv. bencoding. Pro převod do lidsky čitelného textu existují nástroje, včetně těch dostupných online[3]. Obrázek 2 ukazuje příklad dekódovaného souboru typu .torrent. Tracker je specifikovaný položkou *announce* a infohaš v položce *info hash*.

```
            name: debian-9.6.0-amd64-netinst.iso
        filename: debian-9.6.0-amd64-netinst.iso.torrent
         comment: "Debian CD from cdimage.debian.org"
            date: 10.11.18 04.17.25 AM (1541852245)
      created by: PHP BitTorrent
           files: debian-9.6.0-amd64-netinst.iso

            size: 297984.00 kb
        announce: http://bttracker.debian.org:6969/announce
   announce list:

       info hash: f71e7defc014563fc7d8ffe26f759b2518c30f34
```

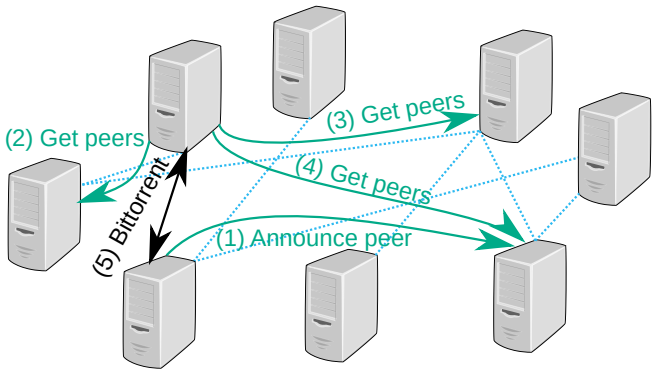
Obrázek 2. Ukázka obsahu souboru debian-9.6.0-amd64-netinst.iso.torrent, viz https://cdimage.debian.org/debian-cd/current/amd64/bt-cd/

Trackery dělíme na veřejné a privátní. *Veřejné trackery* nabízejí zprostředkování peerů každému, zatímco *privátní trackery* jsou dostupné jen úzké komunitě uživatelů. Privátní trackery typicky fungují na pozvání a snaží se zajistit, aby jednotlivý uživatelé nabízeli obsah v míře odpovídající jejich stahování dat od ostatních. Proto jsou jednotlivý uživatelé identifikování pomocí unikátního ID, které je v souboru .torrent uvedeno jako součást URL trackeru. V případě, že privátní tracker poskytuje webové stránky, může si uživatel prohlédnout svá dřívější stahování, množství dat poskytnutých ostatním, množství stažených dat apod.

Později byla nutnost komunikace s trackerem doplněna o možnost využití *distribuované hešovací tabulky* (DHT) [2, 7, 8, 11]. Obrázek 3 ukazuje fungování DHT. (1) Nejdříve počítač A zašle zprávou *Announce peer* informaci o tom, že poskytuje data identifikovaná konkrétním infohašem. (2) Počítač B projeví zájem o data identifikovaná stejným infohašem, ale nezná identitu žádného počítače nabízejícího daná data. Pokud však zná nejméně jeden uzel zapojený do DHT (např. dobře známé uzly `router.bittorrent.com`, `dht.transmissionbt.com`, či `router.utorrent.com`), může se postupným doptáváním uzlů zprávami *Get peers* (2)-(4) dostat, až k uzlu C, který má informaci o tom, že počítač A na-

[3] http://beautifytools.com/torrent-decoder.php

bízí poptávaná data. Vyhledávání v tabulce DHT funguje na základě ID uzlů a je zaručeno, že se každým dotazem poptávající přiblíží poptávanému obsahu. (5) Jakmile počítač B zjistí, že počítač A nabízí poptávaná data, může se mezi A a B návazat obvyklá komunikace mezi peery tvořící swarm protokolu BitTorrent [3, 9].



Obrázek 3. S využitím DHT je možné nalézt účastníky swarmu bez existence centrálního uzlu.

URI využívající schéma *magnet* [6] usnadňuje uživatelům výměnu dat s pomocí DHT, protože může nést požadovaný infohaš, např. URI `magnet:?xt=urn:btih:f71e7defc014563fc7d8ffe26f759b2518c30f34` poskytuje pro stahování dat obdobnou informacemi jako soubor .bittorrent uvedený na obrázku 2. V případě souboru .torrent je pro zjištění peerů nutné kontaktovat centrální tracker, při použití DHT je jak nalezení peerů, tak výměna dat decentralizovaná.

3 Případy použití sítě BitTorrent

Tato sekce obsahuje případy použití a monitorování sítě BitTorrent z pohledu různých aktérů, od obyčejných uživatelů přes policejní složky po ochranu duševního vlastnictví se zaměřením na bezpečnostní výzkum. Každá z následujících podsekcí popisuje jeden případ užití.

3.1 Obyčejný uživatel sítě BitTorrent

Tento případ užití se zabývá obvyklým použitím sítě Bittorent, tzn. jejím klasickým uživatelem. Uživatel si chce stáhnout nějaký obsah, je možné, že na internetu narazil na pro něj zajímavý odkaz schéma magnet [6], či dostal doporučení na zajímavý film, nebo se pokouší stáhnout nový obraz své oblíbené distribuce GNU/Linux.

V případě, že má k dispozici URI schématu magnet, může jej zadat do svého klienta, ten se připojí k DHT, najde peery nabízející požadovaný obsah, naváže s nimi spojení a začne stahovat. Po dokončení stahování jednotlivých dílů je začne nabízet ostatním připojeným peerům. Po dokončení stahování se uživatel často odpojí ze swarmu.

Obdobným způsobem postupuje uživatel mající k dipozici soubor formátu .torrent. Interakce s klientem sítě BitTorrent je pro něj stejná. Jediný rozdíl spočívá v tom, že klient nehledá peery ve swarmu pomocí DHT, ale naváže spojení s trackerem.

V případě, že uživatel nemá k dispozici ani odkaz ve formě URI schématu magnet ani soubor .torrent, musí si jedno opatřit vlastní cestou. Typicky navštěvuje indexační stránku jako je např. https://thepiratebay.org/, může však zvolit variantu vyhledávání na různých webech, které umožňují vkládat uživatelský obsah jako jsou pastovací služby[4], či internetová fóra[5].

Pokud uživatel využívá služeb privátního trackeru, ten bývá vybaven také indexační službou a nabízí uživateli soubory formátu .torrent, či odpovídající linky schématu URI.

Ve všech případech je typickému uživateli sítě Bittorent stažením souboru splněn jeho zájem. Takový uživatel tedy nemá zájem o poznávání fungování sítě BitTorrent, vylepšování nástrojů, či monitorování a není tedy kandidátem na využívání výsledků projektu *Integrovaná platforma pro zpracování digitálních dat z bezpečnostních incidentů*. Tomuto případu využití se tedy již dále nebudeme věnovat.

3.2 Zjištění infohaše, souborů .torrent a trackerů

Jak bylo zmíněno v sekci 2, znalost infohaše je základním předpokladem pro kontaktování odpovídajícího swarmu. Protože není možné pro konkrétní data jednoznačně odvodit odpovídající infohaš (viz sekci 2) a v následující případech užití je potřeba pracovat s jinými identifikátory dat, např. jménem, je potřeba nalézt způsob, jakým pro konkrétní jméno souboru hledat infohaš.

Soubory .torrent jsou vhodným zdrojem dat pro monitorování nejen proto, že obsahují infohaše, ale také proto, že obsahují údaje o trackerech.

Bakalářská práce Martina Grnáče [5] se mimo jiné zabývala vyhledáváním a analýzou souborů .torrent. V rámci práce vznikl nástroj pro stahování souborů za pomocí vyhledávače Google[6]. V rámci projektu *Integrovaná platforma pro zpracování digitálních dat z bezpečnostních incidentů* jsme se případu užití rozhodli věnovat i dále a pracujeme na indexačním nástroji (především formou diplomové práce Tomáše Kocmana), který dokáže na webových stránkách rozpoznat infohaše a nabídne vyšetřovatelům možnost vyhledávání infohašů patřící k zadanému dotazu. Dokončení nástroje předpokládáme v roce 2019.

[4] Např. https://pastebin.com/search?q=magnet+btih, či https://pastebin.com/search?q=.torrent.

[5] Např. http://www.orangepi.org/orangepibbsen/forum.php?mod=viewthread&tid=3208.

[6] https://www.google.com

Diskuzí s příslušníky Policie ČR bylo zjištěno, že průběžná indexace není potřeba. Poptávaná je funkcionalita zaměřená na cílené vyšetřování. Tedy například v rámci vyšetřování je potřeba odhalit trackery nabízející obsah identifikovaný názvem jako je `mall\.?cz(-accounts)?(-emails)?-passwords-zip` a volitelně název webové služby, např. https://pastebin.com. Cílem nástroje pak bude indexovat konkrétní stránky obsahující řetězec odpovídající regulárnímu výrazu a typem vyhledávaných dat, např. soubor .torrent, link ve schématu magnet apod.

3.3 Kdo sdílí moje data?

V případě ukradených dat, jako byla např. databáze hesel uniklých ze serveru MAll.cz, pro ochranu autorských práv apod. je potřebné, či vhodné monitorovat šíření dat. Monitorování stahovaných dat může být výhodné i v případě legálního sdílení např. obrazů distribuce GNU/Linux pro vytváření statistik.

Takové monitorování se může sestávat z hledání infohašů a následného prohledávání DHT, či z vyhledávání trackerů v souborech .torrent a následného získávání informací o členech swarmu az pomocí trackeru. Prerekvizitou je tedy scénář použití popsaný v sekci 3.2.

Monitorováním sítě BitTorrent se zabývají diplomová práce Davida Bezděka [1] a bakalářská práce Martina Vaška [10] obhájené v roce 2018. Tyto práce vznikly jako součást projektu *Integrovaná platforma pro zpracování digitálních dat z bezpečnostních incidentů*. Nástroj dht-crawler[7] vzniklý v roce 2018 adresuje právě tento případ užití. Nástroj dokáže na základě vstupu ve formě infohaše, URI schématu magnet, či souboru .torrent prohledat DHT a vypsat všechny známé uzly, které se podílejí na sdílení odpovídajících dat.

3.4 Detekce seedboxů

Seedbox je stroj, který slouží především k nabízení dat ostatním členům sítě BitTorrent. Seedbox je dlouhodobě připojen do všech swarmů a nabízí lokálně uložená data ostatním. Seedboxy se v síti BitTorrent využívají pro zvýšení uploadů konkrétního uživatele, či pro zvýšení dostupnosti konkrétních dat. Jejich vlastnosti je převaha uploadu nad downloadem. Jejich přítomnost v síti jí může nadměrně zatěžovat a proto odhalení seedboxů může být přínosné pro správce sítě.

Na tento případ užití se zaměřila bakalářská práce Martina Grnáče [5] úspěšně obhájená v roce 2018 řešená jako součást projektu *Integrovaná platforma pro zpracování digitálních dat z bezpečnostních incidentů*. Práce využívá statistiky využití sítě protokolu NetFlow.

[7] https://www.fit.vutbr.cz/~polcak/prods.php?id=581, zdrojové kódy dostupné na https://github.com/polcak/DHT-crawler

3.5 Analýza protokolu BitTorrent

Výzkumníci, studenti a vývojáři se zabývají detaily fungování protokolu BitTorrent, jako jsou jednotlivé zprávy, návaznosti, objevování nových peerů, strategie stahování dílů apod.

Vyšetřovatel mající k dispozici zachycenou síťovou komunikaci ve formátu pcap používá nástroj Wireshark[8]. Wireshark má sice disektory schopné BitTorrent a jeho provoz analyzovat a pársovat [12], nicméně je nutné disektory aktivovat manuálně.

Bakalářská práce Daniela Florka řešená v rámci projektu *Integrovaná platforma pro zpracování digitálních dat z bezpečnostních incidentů* a úspěšně obhájená v roce 2018 se snaží hledat automatizované přístupy k vyhledávání provozu BitTorrent. Práce se zaměřila na vyhledávání řetězce „GET /announce“ a „GET /scrape“ v komunikaci, čímž je možné najít komunikaci s trackerem, ale např. DHT není v práci řešená.

Bakalářská práce Daniela Florka [4] také obsahuje analýzu podpory šifrované komunikace mezi trackery inzerovanými v souborech .torrent za využití nástroje [5] zmíněného v sekci 3.2. Pouze 1 % trackerů používá šifrování TLS (schéma https), 99 % analyzovaných souborů však obsahuje alespoň jeden tracker, který používá schéma https. Klient, který je nastavený tak, aby nuceně používal TLS tedy v naprosté většině případů najde nejméně jeden tracker. Na druhou stranu klienti, kteří si TLS nevynucují budou často využívat nešifrovanou variantu komunikačního protokolu BitTorrent. Tím pádem se jejich provoz při komunikaci s trackerem přenáší po síti v jednoduše čitelné podobě.

3.6 Výzkumné problémy

Výzkumníci a následně vývojáři a studenti se také zabývají různými výzkumnými problémy spojenými se sítí BitTorrent jako je monitorování víření, či BitTorrent sync[9]. Protože se však nejedná o bezpečnostní výzkum, *Integrovaná platforma pro zpracování digitálních dat z bezpečnostních incidentů* se tomuto případu využití více nevěnuje.

3.7 Analýza minulosti aneb chci vědět, kdo v minulosti ...

Pro vyšetřování trestné činnosti může být podstatné zjišťovat stav sítě BitTorrent. Typicky však data z minulosti již nelze získat. Pro protokol Bittorent neexistuje služba podobná CollecToru[10]. Data jsou navíc distribuovaná a vytváření trackerů nepodléhá jednotnému dohledu.

Pro zkoumání minulosti je možné využít metainformace typu NetFlow. Příkladem může být detekce seedboxů zmiňovaná v podsekci 3.4. V případě využití privátních trackerů je možné při vyšetřování trestné činnosti získat data ukládaná samotným trackerem na základě soudního povolení.

[8] https://www.wireshark.org/
[9] https://www.resilio.com/
[10] https://collector.torproject.org/

4 Závěr

Případy použití sítě BitTorrent v souvislosti s bezpečnostním výzkumem často potřebují znát infohaš, či trackery odpovídající konkrétnímu obsahu. Projekt *Integrovaná platforma pro zpracování digitálních dat z bezpečnostních incidentů* se tímto případem použití zabýval již v roce 2018 a vznikl nástroj pro vyhledávání souborů .torrent [5], v roce 2019 budou práce na tomto případu užití pokračovat a budeme se zabývat nástrojem pro prohledávání webu.

V rámci projektu *Integrovaná platforma pro zpracování digitálních dat z bezpečnostních incidentů* v roce 2018 vznikl nástroj *dht-crawler*, který prohledává DHT [1, 10]. Dále jsme se zabývali automatickou detekci protokolu BitTorrent [4] a detekci seedboxů [5].

Literatura

[1] Bezděk, D.: *Monitorování peerů sdílejících torrenty*. Diplomová práce, Vysoké učení technické v Brně, Fakulta informačních technologií, 2018.
URL http://www.fit.vutbr.cz/study/DP/DP.php?id=20316

[2] Chroboczek, J.: *BitTorrent DHT Extensions for IPv6*. 2009, BitTorrent BEP-32, Standards Track, status draft, verze b0f04e1bfb9f04da2590370c8dcd46b58f35b45b, 2009.
URL http://www.bittorrent.org/beps/bep_0032.html

[3] Cohen, B.: *The BitTorrent Protocol Specification*. 2017, BitTorrent BEP-3, Standard, status final, verze 0e08ddf84d8d3bf101cdf897fc312f2774588c9e, 2017.
URL http://bittorrent.org/beps/bep_0003.html

[4] Florek, D.: *Detekce provozu protokolu BitTorrent*. Bakalářská práce, Vysoké učení technické v Brně, Fakulta informačních technologií, 2018.
URL http://www.fit.vutbr.cz/study/DP/BP.php?id=20313

[5] Grnáč, M.: *Detekce seedboxů v síti BitTorrent*. Bakalářská práce, Vysoké učení technické v Brně, Fakulta informačních technologií, 2018.
URL http://www.fit.vutbr.cz/study/DP/BP.php?id=20465

[6] Hazel, G.; Norberg, A.: *Extension for Peers to Send Metadata Files*. 2017, BitTorrent BEP-9, Standards Track, status accepted, verze cfd7cd53addbea7e87363ae1a52e2e4b397df3a9, 2017.
URL http://bittorrent.org/beps/bep_0009.html

[7] Loewenstern, A.; Norberg, A.: *DHT Protocol*. 2017, BitTorrent BEP-5, Standards Track, status accepted, verze d7976ebbd42e46e3a14a5b7fa50b1cc89a7c568d, 2017.
URL http://bittorrent.org/beps/bep_0005.html

[8] Maymounkov, P.; Mazières, D.: Kademlia: A Peer-to-Peer Information System Based on the XOR Metric. In *Peer-to-Peer Systems*, editace P. Druschel; F. Kaashoek; A. Rowstron, Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, ISBN 978-3-540-45748-0, s. 53–65.

[9] Norberg, A.: *uTorrent transport protocol*. 2017, BitTorrent BEP-29, Standards Track, status accepted, verze 023256c7581a4bed356e47caf8632be2834211bd, 2017.
URL http://bittorrent.org/beps/bep_0029.html

[10] Vaško, M.: *Monitorování peerů BitTorrent na základě informací z distribuované hašovací tabulky*. Bakalářská práce, Vysoké učení technické v Brně, Fakulta informačních technologií, 2018.
URL http://www.fit.vutbr.cz/study/DP/BP.php?id=20527

[11] Wang, L.; Kangasharju, J.: Measuring large-scale distributed systems: case of BitTorrent Mainline DHT. In *IEEE P2P 2013 Proceedings*, 2013, ISSN 2161-3559, s. 1–10.

[12] Wireshark Foundation: BitTorrent — The Wireshark Wiki. [[online]], naposledy navštíveno 12.12.2018.
URL https://wiki.wireshark.org/BitTorrent