VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

avast

Avast – Metody pro extrakci a detekci vzorů v programovém kódu

Stav řešení projektu v roce 2018

1. Úvod

Rok 2018 byl pro projekt společnosti Avast Software druhým rokem řešení. Vzhledem k celkové změně řešitelského týmu došlo také k zásadní změně zaměření. Po společné dohodě se společností Avast byl výzkum a vývoj pro detekci vzorů v programovém kódu přesunut do interního týmu společnosti. Novým cílem tohoto projektu je tak průzkum v oblasti získávání znalostí, tzv. *threat intelligence*, v oblasti šíření škodlivých emailů.

V průběhu roku 2018 byl nasazen SMTP honeypot, pomocí kterého byla sbírána data v podobě spamu. První část řešení projektu se zaměřovala především právě na nasazení honeypotu a následnou spolehlivou analýzu všech získaných dat a uložení získaných informací ve vhodné podobě do perzistentního úložiště. Byly také prozkoumány možnosti shlukování spamu do kampaní a navržen postup pro zpracování dat z honeypotu v reálném čase. Následně byl prototyp tohoto systému implementován. Rovněž byl vytvořen jednoduchý dashboard, informující o získávaných znalostech v reálném čase.

2. Průběh řešení v roce 2018

Řešení probíhalo dle průběžných požadavků společnosti Avast Software. Společné schůzky, na kterých byl diskutován postup, probíhaly každý týden. Ty byly doplňovány dalšími osobními setkáními přímo ve společnosti Avast.

2.1. Získávání znalostí z SMTP honeypotu

Prvním krokem bylo nasazení SMTP honeypotu pro získávání dat. Byl využit blíže nespecifikovaný poskytovatel VPS, na který byl nasazen open-source SMTP server SHIVA[1]. Tento software nám umožňuje zachytit a analyzovat všechny příchozí emaily. Díky němu můžeme využít autentizace, čímž odstraníme problémy s provozem open relay SMTP serveru. Vzhledem k tomu, že server není nikde inzerován k veřejnému využití, procházejí přes něj pouze emaily od uživatelů, kteří se k přihlašovacím údajům dostali neautorizovaně.

[1] Viz https://github.com/shiva-spampot/shiva

Kromě funkčního honeypotu je nutné věnovat se na straně vzdáleného serveru převážně stále otevřeným problémům. Mezi ty patří problematika práce s malware, kvůli čemuž jsme vystaveni stálému tlaku ze strany poskytovatele VPS. Stejně tak problematika distribuce autentizačních údajů a rozšíření celého řešení je v plánu v následujícím období a bude diskutováno v kapitole 4.

Následně byla vyvinuta sada nástrojů, která nám umožňuje automatizovaný sběr dat, a to jak z jednoho serveru, tak celé sítě, tzv. honeynetu. Tyto nástroje za nás pravidelně shromažďují všechna data a volitelně také spouští jejich analýzu, uložení získaných znalostí do databáze a aktualizaci dashboardů.

Pro analýzu samotných dat byl využit modul *email* jazyka Python 3. Byl implementován wrapper, který poskytuje vhodné rozhraní pro následné uložení do databáze. Tento wrapper navíc doplňuje funkcionalitu modulu *email* o výpočet hešů, detekci jazyka, uložení jména souboru a další. Tento nástroj je následně spouštěn paralelně nad sadou zachycených emailů.

Data jsou ukládána do PostgreSQL databáze. Tato databáze byla vybrána z důvodu dobré zkušenosti s SQL databázemi z projektu Honeynet[1], v rámci kterého byl vyvinut zde využívaný nástroj SHIVA. Před samotným uložením získaných informací z EML souborů dojde k detekci duplicit, tedy shlukování do kampaní. To je v aktuálním prototypu řešeno s využitím nástroje *ssdeep*. Ten funguje na principu *Context Triggered Piecewise Hashing*[2]. Zkoumání škálovatelnosti tohoto řešení je opět jeden z cílů v dalším období projektu.

Dále byl proveden průzkum v oblasti aktuálního využití technik *data mining* pro získávání znalostí ze spamu. Byly identifikovány dvě oblasti potenciálního využití těchto technik v našem systému. V prvním případě by bylo možné pokročilé techniky využít při agregaci kampaní. Zmíněný ssdeep nemusí být sám o sobě dostatečně přesný a potýká se se špatnou škálovatelností. Z používaných technik, které jsou blíže popsané v interním článku shrnující naši práci, viz následující kapitolu 3, by bylo zajímavé využít strukturu FP-stromu pro shlukování spamových kampaní.

Druhou oblastí využití data mining technik je získávání dodatečných znalostí ze získaného data setu. Příkladem je technika dolovaní dat z textu nebo vytváření asociačních pravidel. K tomu by bylo možné využít právě zmíněného FP-stromu.

V neposlední řade byly implementovány konektory k dalším systémům společnosti Avast Software. Mezi ty patří automatické zasílání škodlivých emailů a jejich příloh do centrálního úložiště malware a hlášení o škodlivých URL odkazech obsažených v tělech emailů.

S využitím protokolu MQTT bylo dále implementováno zasílání zpráv v reálném čase, které umožňuje vizualizaci těchto dat. Tento systém je opět ve stavu prototypu a informuje uživatele pouze o kvantitativních vlastnostech spamu, případně o jeho původu. V budoucnosti bude systém napojen na informace z našeho analyzátoru, což umožní posunout vizualizaci na další úroveň.

[1] Viz https://www.honeynet.org/

[2] Viz https://www.sciencedirect.com/science/article/pii/S1742287606000764

Zároveň byla provedena statistická analýza data setu, který byl nasbírán v průběhu celého roku 2018. Byla sledována země původu spamu, způsob klamání uživatelů, typy příloh nebo jazyk těla emailů. Výsledky byly shrnuty v interní závěrečné zprávě, viz následující kapitolu.

3. Výstupy řešení

Výsledkem práce v roce 2018 je prototyp systému pro analýzu zachyceného spamu v reálném čase a návrh vizualizace získaných znalostí. Tento nástroj má k dispozici společnost Avast Software. Prototyp také komunikuje s dalšími komponentami oddělení TheatLabs a přispívá tak celkově k efektivnějšímu boji proti kybernetickým hrozbám.

Dalším výstupem je interní článek, popisující výzvy při získávání threat intelligence v oblasti SMTP honeypotů. V článku, který je k dispozici společnosti Avast Software, jsou diskutována řešení použitá při implementaci prototypu. Jsou zde také shrnuty získané znalosti z nasbíraných dat v průběhu roku 2018.

4. Plány na další období

Práci v roce 2019 je možné rozdělit na tři části. Jedná se o zlepšování funkcionality na straně honeypotů, na straně zpracování dat a vizualizace. Všechny tyto části jsou aktuálně ve stavu prototypu a je možné je dále zlepšovat.

4.1. Rozšíření honeypotů

Jedním z hlavních úkolů v roce 2019 bude rozšíření honeypotu v globálním měřítku. To nám poskytne nejen více dat, ale také zvýší objektivitu získaného pohledu na hrozby, které se šíří. A to je právě cílem tohoto projektu. Zároveň bude nutné sledovat škálovatelnost na straně backendu, tedy schopnost celého systému vyrovnat se zvyšujícím se množstvím dat z více zdrojů.

Dalším cílem je úprava honeypot software tak, aby byl co nejméně výpočetně náročný a běžel i na nevýkonném serveru. Aktuální funkčnost nástroje SHIVA je pro naše účely příliš složitá a zbytečně spotřebovává zdroje VPS.

4.2. Zvýšení spolehlivosti zpracování

Jak již bylo zmíněno, s rozšiřováním sítě honeypotů bude nutné monitorovat funkčnost analyzujících nástrojů a škálovatelnost databáze. To souvisí s tvorbou testovací sady, která bude zajišťovat integritu celého řešení při provedených úpravách.

Dále je v plánu implementovat pokročilé metody data mining, které nám poskytnout dodatečné znalosti z extrahovaných dat.

4.3. Pokročilá vizualizace

Aktuální řešení vizualizace je velmi jednoduché a nevyužívá potenciál všech informací uložených v databázi. Chtěli bychom rozšířit naši MQTT infrastrukturu tak, aby v budoucnu zvládala pojmout plánované rozšíření sítě, a to v rozsahu vizualizace všech analyzovaných informací. To bude zahrnovat také rozšíření analyzujících skriptů, které budou získané informace ihned odesílat.