Os meus dados de fato vazaram? Uma análise de serviços que monitoram vazamentos de dados na Internet

Ariel Góes de Castro, Felipe Antunes Quirino,
Francisco Germano Vogt, João Otávio Chervinski[1], Diego Kreutz[1]

[1]Laboratório de Estudos Avançados (LEA)
Universidade Federal do Pampa (UNIPAMPA)

{agdecastro, felipeantunesquirino, franciscogermanovogt,

joaootavio}@alunos.unipampa.edu.br, kreutz@unipampa.edu.br

Resumo. *O número de usuários de serviços online não para de crescer. A maioria destes serviços armazena informações sigilosas dos usuários, como credenciais de acesso (login e senha), dados de cartões de crédito e outras informações pessoais. Devido a isto, estes serviços tornaram-se alvos de cibercriminosos. O resultado é um número e volume crescentes de vazamentos de dados sigilosos. Para verificar se seus dados foram vazados na Internet, usuários podem recorrer a serviços online especializados, como o Avast Hack Check e o Have I Been Pwned, que coletam dados de vazamentos ocorridos. O objetivo deste trabalho é analisar os dados destes dois serviços e verificar a consistência e o nível de confiabilidade das informações apresentadas aos usuários. Os resultados iniciais indicam que aproximadamente 50% das credencias vazadas, não cifradas, estão corretas de acordo com os respectivos usuários.*

1. Introdução

A segurança e a privacidade no armazenamento de informações sigilosas, como email, credenciais de acesso, CPF e outros dados privados, é de interesse da maioria dos usuários de serviços online. Ao fornecer este tipo de informação para uma empresa, o usuário deposita a sua confiança nos sistemas da entidade. Consequentemente, a segurança das informações armazenadas deveria ser uma das maiores prioridades das empresas, visto que um vazamento de dados prejudica sua reputação, pode causar a redução de seu valor de mercado [Neate 2018, Machado et al. 2019] e até mesmo levar a pedido de proteção contra falência, como foi o caso recente da empresa AMCA nos EUA [Osborne 2019].

Infelizmente, os dados dos usuários nem sempre são tratados e armazenados levando em consideração aspectos de segurança e privacidade. Em 2019, mais de 2,7 bilhões de endereços de email e senhas foram divulgados publicamente de uma única vez, em um vazamento que foi denominado de "Collection #1" [Hunt 2019]. Este incidente, somado a muitos outros como os vazamentos de dados da Yahoo [Huang et al. 2018], do Facebook [Turner 2019], da British Airways [News 2019], da Quest Diagnostics [McKay 2019] e de várias outras empresas [Machado et al. 2019], expõe as dados sensíveis dos usuários, colocando-o em risco uma vez que agentes maliciosos frequentemente utilizam estes dados para realizar golpes e novos roubos de informação.

Devido ao crescente número de incidentes de segurança [Machado et al. 2019], estão sendo criadas ferramentas e modelos, como o WCGM [Lu et al. 2018], que utilizam

grafos com pesos e técnicas de aprendizado de máquinas para classificar os vazamentos de dados de forma automática. Entretanto, no caso do WCGM, apesar de conseguir analisar uma grande quantidade de dados, o modelo apresenta um grande número de falsos positivos. Métodos específicos de detecção de vazamentos também tem sido propostos na literatura, como os algoritmos para reconhecer padrões em dados e detectar o vazamento de informações sigilosas com um baixo número de falso positivos [Shu et al. 2015]. Dois algoritmos, um de amostragem e um de alinhamento de sequências, são capazes de identificar segmentos de informações vazadas e detectar até mesmo vazamentos parciais de dados. Entretanto, nenhum dos trabalhos encontrados na literatura investiga a qualidade e corretude dos dados de plataformas online de consulta de dados vazados.

As plataformas online Have I Been Pwned e Avast Hack Check foram desenvolvidas para permitir aos usuários, de uma forma simples e rápida, verificar se seus endereços de email (e respectivos dados pessoais como credenciais) estão contidos em bases de dados que monitoram e agregam informações de vazamentos de dados. A Avast Hack Check notifica por email os usuários quando suas senhas, utilizadas nos mais diversos tipos de sites e sistemas online, são vazadas. Entretanto, um dos desafios destas plataformas está relacionado à qualidade e a confiabilidade dos dados, isto é, determinar se o vazamento de fato ocorreu e se os dados estão corretos.

Este trabalho tem por objetivo realizar um estudo inicial da qualidade e confiabilidade dos resultados apresentados nas buscas das plataformas Have I Been Pwned e Avast Hack Check. A pesquisa pode ser dividida em três etapas. Primeiro, foi realizado um levantamento e análise de dados, utilizando um conjunto limitado e conhecido de usuários, para verificar se as senhas informadas como vazadas já foram, de fato, utilizadas pelos respectivos usuários. Segundo, as estatísticas das duas plataformas foram analisadas a fim de encontrar similaridades. Finalmente, foram realizados também testes utilizando uma lista pública de 50.664 emails encontrados na Web para verificar se as estatísticas dos vazamentos é similar a distribuição observada nos casos anteriores.

2. Metodologia

Coleta dos dados. As informações reunidas para a análise foram coletadas a partir de três fontes distintas. Os conjuntos de dados são compostos por endereços de email de usuários, que são a informação utilizada pelas ferramentas analisadas para verificar a ocorrência de vazamentos de dados. Os três processos de coleta são descritos a seguir.

Conjunto de Dados I: A primeira coleta de dados consistiu na aquisição de dados de 50.664 endereços de email disponibilizados na Internet. Este conjunto de dados é composto por endereços de email cadastrados em um serviço de *marketing* online. Esta coleção de endereços favorece a visualização da distribuição dos vazamentos considerando que ela contém informações de usuários (emails) que acessam vários serviços distintos.

Conjunto de Dados II: A segunda coleta consiste em um conjunto de 108 endereços de email conhecidos, obtidos a partir de uma lista de contatos de uso pessoal. Isto garante que são emails efetivamente utilizados. Este conjunto de dados permite comparar os resultados estatísticos das plataformas entre uma lista de emails conhecidos e uma de endereços desconhecidos (Conjunto de Dados I).

Conjunto de Dados III: A terceira coleta de dados foi realizada com a ajuda de vo-

luntários. Primeiro, foi criado e disponibilizado um formulário online, utilizando o serviço Google Forms, para coletar informações sobre os vazamentos de dados dos usuários. Segundo, cada participante acessa a plataforma Avast Hack Check para verificar se seus dados foram vazados. A Avast Hack Check envia um email com os detalhes dos vazamentos, incluindo sites, senhas e outros detalhes. Terceiro, o participante informa no formulário online se as senhas identificadas pela plataforma estão corretas, isto é, são ou já foram utilizadas por ele.

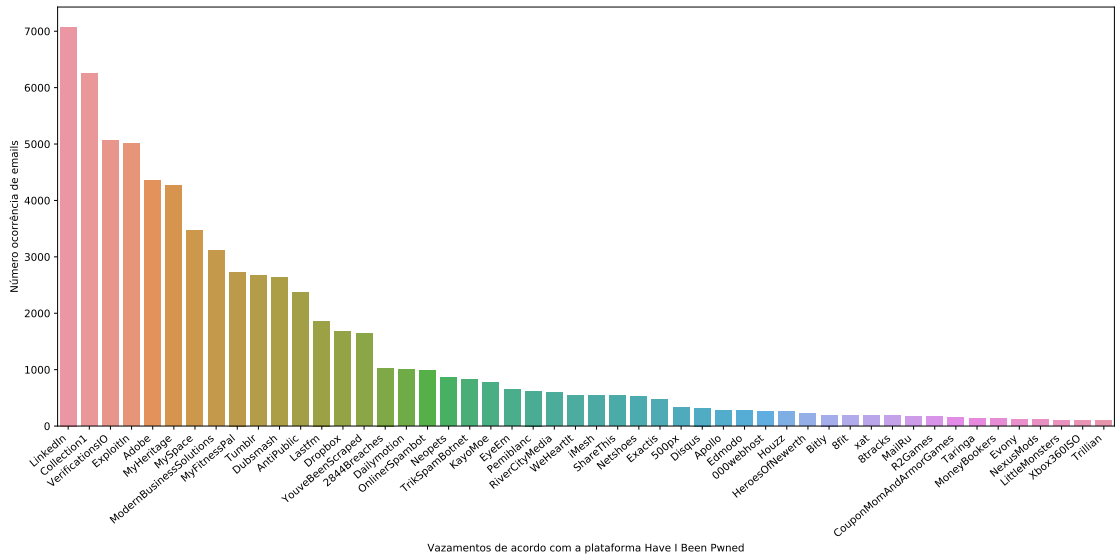
Processamento e análise dos dados. Para automatizar o processo de verificação dos endereços de email dos três conjuntos de dados, foram criados dois scripts. Os scripts preenchem automaticamente os formulários online da plataforma Have I Been Pwned, coletam os resultados e geram os gráficos estatísticos correspondentes (ver detalhes dos gráficos e análise na Seção 3).

3. Resultados

3.1. Estatísticas do Conjunto de Dados I

As Figuras 1 e 2 apresentam, respectivamente, os 50 e os 10 sites com maior quantidade de dados vazados levando em consideração os 50.664 emails do primeiro conjunto de dados. Como pode ser observado, sites conhecidos e amplamente utilizados, como LinkedIn, Adobe, MySpace e Tumblr, estão entre os 10 com o maior número de dados vazados. Segundo estas estatísticas, para este conjunto de dados, o LinkedIn, sozinho, vazou dados de cerca de 13,5% dos emails testados.

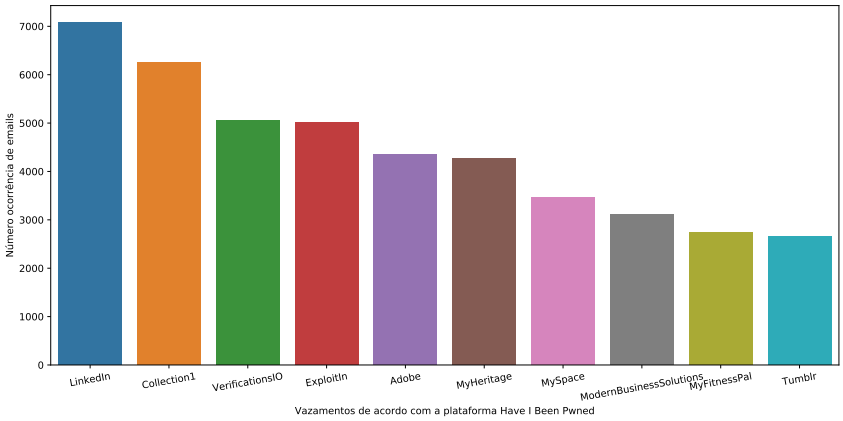
Figura 1. 50 maiores vazamentos do conjunto de 50.664 e-mails



3.2. Estatísticas do Conjunto de Dados II

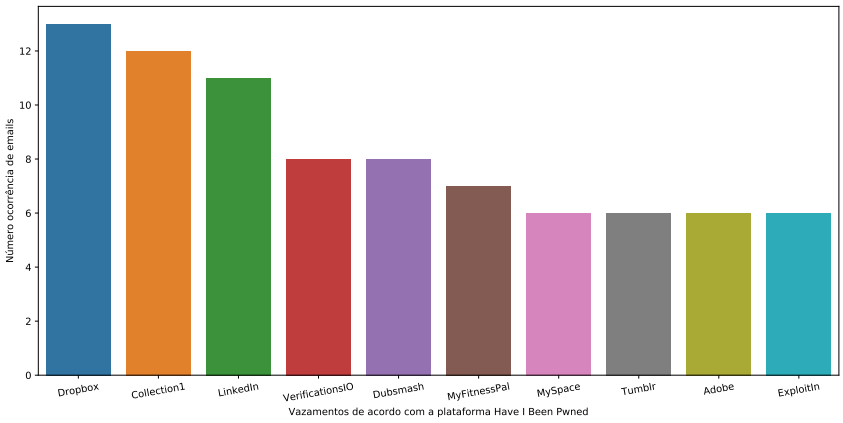
A Figura 3 apresenta as 10 principais fontes de vazamento de dados para o segundo conjunto de dados (e-mails de listas de contatos dos autores do paper). Como pode ser observado, o Dropbox representa a maior fonte de vazamento, com 13 e-mails. Vale ressaltar que, assim como no primeiro conjunto de dados, os sites LinkedIn, Adobe, MySpace,

Figura 2. 10 maiores vazamentos do conjunto de 50.664 e-mails



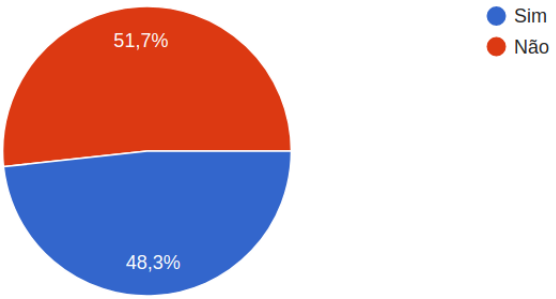
Tumblr, entre outros, aparecem novamente entre os 10 com o maior número de dados vazados. Esta similaridade aponta para uma tendência, que aparentemente independe do conjunto de dados de entrada (emails) utilizado para alimentar as duas plataformas.

Figura 3. 10 maiores vazamentos do conjunto de 108 e-mails.



3.3. Estatísticas do Conjunto de Dados III

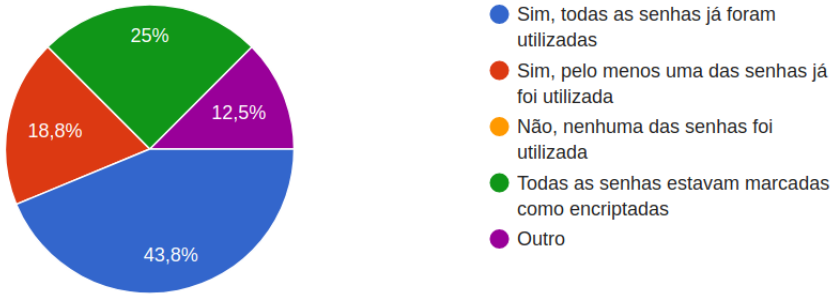
Figura 4. Porcentagem de usuários com credenciais vazadas.



Os dados desta seção correspondem a estatísticas e análises realizadas na plataforma Avast Hack Check. A Figura 4 resume o resultado da pesquisa online com os

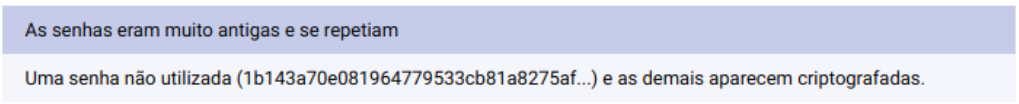
usuários. Como pode ser observado, praticamente 50% dos usuários afirmaram que suas credenciais já foram, ao menos uma vez, vazadas segundo os dados apresentados pela plataforma Avast Hack Check. A porcentagem de confirmações é alta e assegura que, de fato, há dados verídicos e corretos na plataforma. Entretanto, obviamente, isto não significa que podemos assumir que os dados são todos de qualidade e confiáveis. Sabemos que é relativamente comum existir lixo (e.g. dados involuntariamente corrompidos, dados voluntariamente corrompidos, dados forjados) em meio a dados vazados na Internet.

Figura 5. Informações sobre as senhas vazadas



A Figura 5 apresenta dados complementares aos apresentados na Figura 4. Como pode ser observado, 43,8% dos participantes do levantamento de dados informou que todas as senhas apresentadas pela plataforma condiziam com senhas que já haviam sido utilizadas em algum momento. Em segundo lugar, 25% participantes informaram que todas as senhas estavam cifradas (e.g. vazamento de senha cifrada do Dropbox). Consequentemente, não foi possível verificar se as senhas eram, ou não, corretas. Em terceiro lugar, 18,8% dos entrevistados confirmaram que pelo menos uma das senhas apresentadas foi, de fato, utilizada. Consequentemente, do total de participantes, 62,6% (43,8% + 18,8%) confirmaram que pelo menos uma de suas senhas já foi vazada. Finalmente, 12,5% informaram outra situação (e.g. nenhuma senha foi apresentada pela plataforma, as senhas eram muita antigas e se repetiam). A Figura 6 apresenta dois exemplos de feedback dos participantes.

Figura 6. Relatos sobre os vazamentos reportados

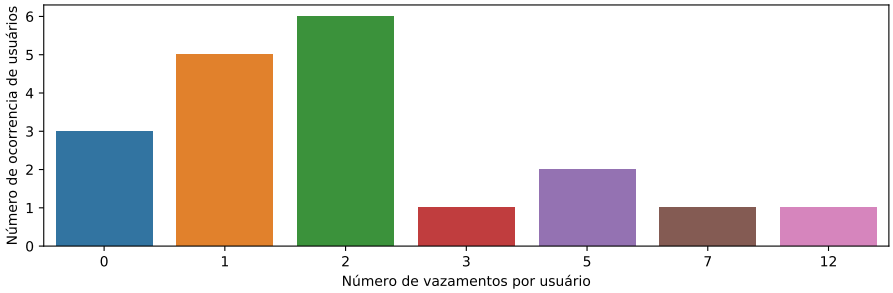


A Figura 7 resume o número de contas comprometidas. Cada conta corresponde a um site/sistema distinto. Enquanto que a maioria dos participantes teve de 0 a 2 contas comprometidas, dois tiveram 7 e 12, respectivamente. Isto reforça a tendência apresentada no gráfico da Figura 1, onde o número de vazamentos é concentrado em aproximadamente 15 sites. Para os demais sites, o número de vazamentos cai rapidamente.

4. Conclusão

Com base nos resultados apresentados, pode-se concluir que a distribuição de vazamentos das plataformas, obtida nos testes com 50.664 emails não verificados, é similar a

Figura 7. Número de cadastros vazados



distribuição obtida com a lista de 108 e-mails verificados. Isso é um indicio de que os resultados dos testes correspondem a realidade.

Segundo as estatísticas apresentadas, 62,6% dos usuários já utilizaram pelo menos uma das senhas vazadas. Isto demonstra que boa parte dos dados apresentados pelas plataformas Have I Been Pwned e a Avast Hack Check correspondem a realidade.

Como trabalhos futuros, podem ser listados: (a) realizar um levantamento de dados mais extensivo, isto é, com um grupo maior de participantes e com questionários mais detalhados; (b) analisar bases de emails maiores; (c) identificar estatisticamente se os emails possuem, de fato, contas nos sistemas indicados (e.g. LinkedIn, MySpace, Dropbox); (d) identificar a porcentagem de falsos positivos das plataformas; (e) cruzar dados com outras fontes, como relatórios técnicos de vazamentos de dados; e (f) avaliar outras plataformas, como a `https://monitor.firefox.com`.

Referências

Huang, X., Lu, Y., Li, D., and Ma, M. (2018). A novel mechanism for fast detection of transformed data leakage. *IEEE Access*, 6:35926–35936.

Hunt, T. (2019). Have i been pwned? `https://haveibeenpwned.com`.

Lu, Y., Huang, X., Ma, Y., and Ma, M. (2018). A weighted context graph model for fast data leak detection. In *IEEE Int. Conf. on Communications (ICC)*, pages 1–6. IEEE.

Machado, R. B., Kreutz, D., Paz, G., and Rodrigues, G. (2019). Vazamentos de Dados: Histórico, Impacto Socioeconômico e as Novas Leis de Proteção de Dados. In *4o WRSeg*. SBC. `http://tiny.cc/wrseg19-dl`.

McKay, T. (2019). Lab Testing Giant Quest Diagnostics Says Data Breach May Have Hit Nearly 12 Million Patients. `http://bit.do/e25Ps`.

Neate, R. (2018). Over $119 billions wiped off Facebook's market cap after growth shock. `http://bit.do/e3dwM`.

News, B. (2019). British Airways faces record £183m fine for data breach. `http://bit.do/e25Q7`.

Osborne, C. (2019). Data breach forces medical debt collector AMCA to file for bankruptcy protection. `http://bit.do/e25Px`.

Shu, X., Zhang, J., Yao, D. D., and Feng, W.-C. (2015). Fast detection of transformed data leaks. *IEEE Transactions on Information Forensics and Security*, 11(3):528–542.

Turner, S. (2019). 2019 data breachers - the worst so far. `http://bit.do/e25MP`.