Universidade Federal de Pernambuco
Centro de Informática
Graduação em Ciência da Computação

O uso de métodos de detecção de outliers na identificação de fraudes em anúncios mobile

Trabalho de Graduação

Aluna: Maria Júlia Godoy Ferreira Lima
Orientador: Ricardo Bastos C. Prudêncio

Recife
Dezembro de 2017

Universidade Federal de Pernambuco
Centro de Informática

O uso de métodos de detecção de outliers na identificação de fraudes em anúncios mobile

Trabalho apresentado ao Programa de Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Aluna: Maria Júlia Godoy Ferreira Lima
(mjgfl@cin.ufpe.br)

Orientador: Ricardo Bastos Cavalcante Prudêncio
(rbcp@cin.ufpe.br)

Recife
Dezembro de 2017

Agradecimentos

Primeiramente, gostaria de agradecer aos meus pais e irmão, por serem exemplos de amor incondicional, e por sempre terem me apoiado, incentivado e inspirado ao longo de todos esses anos. Agradeço aos meus pais por sempre terem estimulado eu e o meu irmão a sermos pessoas estudiosas e dedicadas, e pelo esforço para prover para nós a melhor educação possível. Também agradeço a eles pela dedicação e compromisso na nossa formação como pessoa, nos ensinando bons valores e caráter. À minha mãe, por sempre me apoiar nos momentos de dificuldade e por aguentar meus dramas em semanas de provas e projetos. Também agradeço a minha avó materna (in memorian) por sempre me tranquilizar antes das provas, e por sempre se orgulhar das minhas conquistas. Obrigada por tudo, a minha formação ao longo da vida não teria sido a mesma sem vocês, eu sou muito grata de ter a família que tenho.

Aos demais familiares, por sempre se preocuparem em manter a união da família e por proporcionarem momentos de descontração e bons domingos em Aldeia.

Às amigas que fiz no colégio, por sempre estarem presentes e dispostas a me ouvir, independente do momento. Vocês sempre me incentivaram a ser uma pessoa melhor e compartilharam momentos bons e ruins comigo. Obrigada por todo carinho, cuidado e pelo apoio incondicional de vocês.

Ao professor Ricardo Prudêncio, que foi meu orientador neste trabalho, por ser um excelente professor, e ter dado ótimas sugestões durante o desenvolvimento deste trabalho. Obrigada pela disponibilidade, pela ajuda e pela orientação.

Ao Corpo Docente e funcionários do Centro de Informática, por construírem um Centro de Excelência, que oferece inúmeras oportunidades acadêmicas aos estudantes, e que também fomenta atividades extracurriculares que promovem a integração da comunidade do CIn e a formação pessoal dos estudantes.

Aos meus colegas de turma, por sempre proporcionarem um ambiente de descontração e colaboração durante esses anos. Passamos por muitos momentos difíceis durante a graduação, mas sempre conseguimos nos ajudar e fazer possível a sobrevivência nos finais de período.

Agradeço também aos membros do Clique, que em meio a esse processo, tornaram-se amigos que vou levar para a vida toda. Camila, Simone e Victor, muito obrigada por todos os projetos em grupo, todas as dificuldades ultrapassadas, todas as noites mal dormidas, todos os bons momentos vividos, todas as conquistas compartilhadas, e todo o apoio que vocês me deram na minha vida acadêmica e pessoal. Esses anos teriam sido muito mais difíceis sem a companhia diária de vocês.

Ao PET informática e ao Professor Tutor Fernando Fonseca, por ter me proporcionado a oportunidade de participar em projetos relevantes de ensino, pesquisa, e extensão, sou muito grata por ter impactado tantas pessoas com esses projetos. O PET também me deu a oportunidade de conhecer pessoas incríveis. Eu aprendi muito com todos

os PETianos com quem convivi, e alguns deles se tornaram amigos que vou levar para a vida. A sala do PET se tornou um ambiente fundamental para a minha graduação, vivi lá muitos momentos de construção e aprendizado, e também de fortalecimento de amizades e de descontração. Sou muito orgulhosa de ter feito parte do PET Informática, e é uma parte da graduação que vai deixar muitas saudades.

À In Loco, empresa onde fiz meu primeiro estágio, e onde trabalho atualmente. Obrigada a todos que me proporcionaram aprendizados, e que me ajudaram na decisão do tema e na elaboração deste trabalho. Sinto orgulho de estar colaborando diariamente na construção dessa empresa.

Ao BEPiD (atual Apple Developer Academy) e a todos que constroem esse programa, pela oportunidade de participar de um projeto diferenciado e inovador, sou grata pelos aprendizados, momentos e amigos que fiz no ano de 2015.

Também gostaria de agradecer à Deus, por ter me dado forças nos momentos difíceis e por ter nascido em uma parcela da população que tem melhores condições que os demais. Em um país tão desigual, é um privilégio ter tido as oportunidades que eu tive, e sou muito grata por isso.

At the end of the day, we can endure
much more than we think we can.

— FRIDA KAHLO

Resumo

Nos dias de hoje as fraudes estão presentes em diversos âmbitos, e frequentemente essas fraudes resultam em prejuízos significativos para os diversos negócios afetados. Um setor que vem enfrentando o problema em questão é o de anúncios *mobile*. Estes anúncios, presentes em *smartphones* e *tablets*, podem sofrer diversos tipos de fraude. Atualmente as redes de anúncios pagam os aplicativos parceiros (*publishers*) com base no número de impressões e de cliques em anúncios, então, torna-se interessante aos donos de aplicativos mal intencionados fraudar cliques e impressões, com o objetivo de obter mais lucro. É muito importante identificar este tipo de fraude, pois elas acarretam em um grande prejuízo para as redes de anúncios. Em busca de identificar possíveis fraudadores em uma rede de anúncios foram realizados experimentos utilizando diferentes técnicas para a detecção de *outliers*, pois *outliers*, em geral, podem indicar comportamentos fraudulentos. Os resultados de um dos experimentos não foram bons, mas, com o outro experimento, foi possível apontar alguns usuários e aplicativos como candidatos a fraudadores.

Palavras-chave: anúncios *mobile*, redes de anúncios, fraude, detecção de *outliers*, mineração de dados, estatística

Abstract

Nowadays, fraud is present in many areas, and they often result in significant losses for the various businesses affected. One field that is facing this problem is the one of mobile ads. These ads, present in smartphones and tablets, can experience diverse types of fraud. Ad networks are currently paying for partner applications (publishers) based on the number of impressions and ad clicks, so it is interesting for malicious publishers to cheat on clicks and impressions in order to make more profit. It is very important to identify this type of fraud as they result in a great loss to the ad networks. In order to identify possible fraudsters in an ad network, experiments were conducted using different techniques for the detection of outliers, as outliers in general could indicate fraudulent behavior. The results of one of the experiments were not good, but with the other experiment, it was possible to point out some users and applications as candidates for fraudsters.

Keywords: mobile ads, ad networks, fraud, outlier detection, data mining, statistics

Sumário

1. Introdução

Nos dias de hoje as fraudes estão presentes em diversos âmbitos. Mais especificamente na tecnologia da informação esse tipo de golpe acontece em várias áreas, como o *e-commerce*, telecomunicações, *internet banking*, entre outras [1]. Frequentemente essas fraudes resultam em prejuízos significativos para os diversos negócios afetados, então, é importante buscar maneiras de identificar e evitar esses incidentes.

Um setor que vem enfrentando o problema em questão é o de anúncios *mobile*. Estima-se que em 2013 os anunciantes perderam quase um bilhão de dólares devido a esse tipo de má conduta [2]. De acordo com estudos recentes [3], os anúncios em smartphones e tablets podem sofrer diversos tipos de fraude, que variam desde as relacionadas a cliques até as relacionadas com o percentual de visibilidade do anúncio na tela, por exemplo. Atualmente o sistema funciona de modo que as redes de anúncios *mobile* pagam certas quantias ao proprietário do aplicativo onde o anúncio é exibido, a depender do número de impressões de anúncios e/ou da quantidade de cliques nas propagandas. Então, torna-se interessante aos donos de aplicativos (*publishers*) mal intencionados fraudar cliques e impressões, com o objetivo de obter mais lucro. Portanto, é muito importante detectar esses tipos de fraudes em anúncios *mobile*. Conseguindo identificar *publishers* fraudadores, as redes de anúncios podem se desvincular dos mesmos, evitando prejuízos e mantendo a estabilidade do negócio.

Os estudos recentes sobre detecção de fraudes [1] apontam que quando um conjunto de dados possui observações geradas por fraudes, essas observações geralmente desviam do padrão das observações consideradas normais, então podem ser considerados *outliers* no conjunto de dados. Por este motivo, as técnicas de detecção de *outliers* são amplamente utilizadas com o objetivo de identificar fraudes.

1.1. Objetivos

Este trabalho tem como propósito principal aplicar métodos de detecção de *outliers* em dados reais e não categorizados de anúncios *mobile*, relativos a cliques e impressões. O

objetivo é identificar comportamentos anômalos nesses dados e, dessa maneira, apontar possíveis fraudes.

Os experimentos irão envolver diferentes métodos de detecção de *outliers*, que compreendem técnicas estatísticas baseadas nos valores das observações e na distribuição estatística dos dados. Além disso, serão analisadas diferentes variáveis do conjunto de dados, a fim de utilizar as métricas mais relevantes. Desta forma, será possível fazer uma análise dos resultados, comparar resultados de técnicas diferentes e identificar prováveis usuários ou aplicativos fraudadores.

1.2. Estrutura do Trabalho

Este trabalho está dividido em 5 capítulos, incluindo este que apresenta uma introdução ao tema estudado, além dos objetivos. Os capítulos 2 e 3 contém a fundamentação teórica necessária para o entendimento dos conceitos abordados no trabalho. O capítulo 2 inclui os conceitos sobre anúncios *mobile* e redes de anúncios, enquanto o capítulo 3 é focado em apresentar os conceitos relacionados a fraudes e detecção das mesmas. O capítulo 4 é o principal capítulo do trabalho, pois nele está contido todo o desenvolvimento, incluindo os experimentos realizados e os resultados dos mesmos. O capítulo 5 apresenta a conclusão da pesquisa, além de sugerir trabalhos futuros envolvendo este tema.

2. Anúncios *Mobile*

Os últimos anos aqueceram bastante o mercado de dispositivos móveis, isso se deve ao surgimento e popularização dos *smartphones*. Com a facilidade de acesso a *smartphones*, somado às facilidades que esse tipo de dispositivo provê, esses dispositivos vêm sendo usados pela maior parte da população mundial. Mais especificamente no Brasil, de acordo com uma pesquisa recente da Fundação Getúlio Vargas [4], até o final de 2017 o número de *smartphones* no país será igual ao número de habitantes.

Com essa popularização, o mercado *mobile* tornou-se um segmento que movimenta bilhões ano a ano, tornando-se uma plataforma de negócios, e gerando vários empregos. O crescimento desse mercado fez com que as empresas buscassem maneiras de gerar rendimentos por meio dos aplicativos, ou seja, monetizar essas aplicações.

Existem diversas maneiras de monetizar aplicativos em *smartphones*, entre elas, podemos citar: aplicativos pagos, aplicativos *freemium*, compras dentro do aplicativo, aplicativos por assinaturas, aplicativos patrocinados, e os anúncios, que são o foco desta seção. Nesta seção será apresentado um panorama relativo aos anúncios *mobile*, desde como eles surgiram, passando pelo funcionamento, e mostrando porque esse mercado é tão suscetível a fraudes.

2.1. O que são os Anúncios *Mobile*

Uma pesquisa realizada pela empresa americana *Sweet Pricing* [5] apontou que anúncios dentro do aplicativo é o método de monetização mais popular, com 65% dos aplicativos do estudo exibindo anúncios. Destes, metade utiliza anúncios como única forma de monetização. Com esses dados, é possível entender o quão utilizado esse método é, e, muitas vezes, é a única fonte de renda para os desenvolvedores do aplicativo.

De modo geral, o modelo de anúncios *mobile* funciona de forma de que o desenvolvedor do aplicativo cede um espaço na sua aplicação para veicular anúncios publicitários, em troca de um determinado lucro, que varia de acordo com as métricas

utilizadas e o engajamento do usuário. Uma descrição mais aprofundada acerca de como funcionam as redes de anúncios, e como o lucro final do desenvolvedor é calculado será feita na próxima seção.

Figura 1: Exemplo de anúncio no *app* (Esquerda) e anúncio no *browser* após clique (Direita)

Fonte: *Fun mobility blog* [6]

Esses anúncios possuem diversos formatos, que variam entre *banner*, *interstitial*, anúncios nativos, notificações. Alguns fatores positivos desse modelo são a facilidade de implementação, além do baixo custo, por isso esse modelo vem sendo bastante adotado. Um fator negativo é que é necessário o acesso à internet para viabilizar a veiculação de mídia.

É importante ressaltar que por usar o espaço do aplicativo para anunciar, é preciso ter cautela para não prejudicar a experiência do usuário. Por esse motivo, o desenvolvedor tem a liberdade de escolher tanto o local onde o anúncio será exibido, como também a frequência de exibição, o formato do anúncio, e, dependendo da rede, filtrar os anúncios de acordo com o conteúdo. Dessa maneira, é muito comum encontrar no mercado exemplos de aplicativos que geraram muita receita e engajamento por meio dos anúncios, mas há casos sem sucesso: tudo depende de um bom planejamento, testes com usuários, e do bom senso.

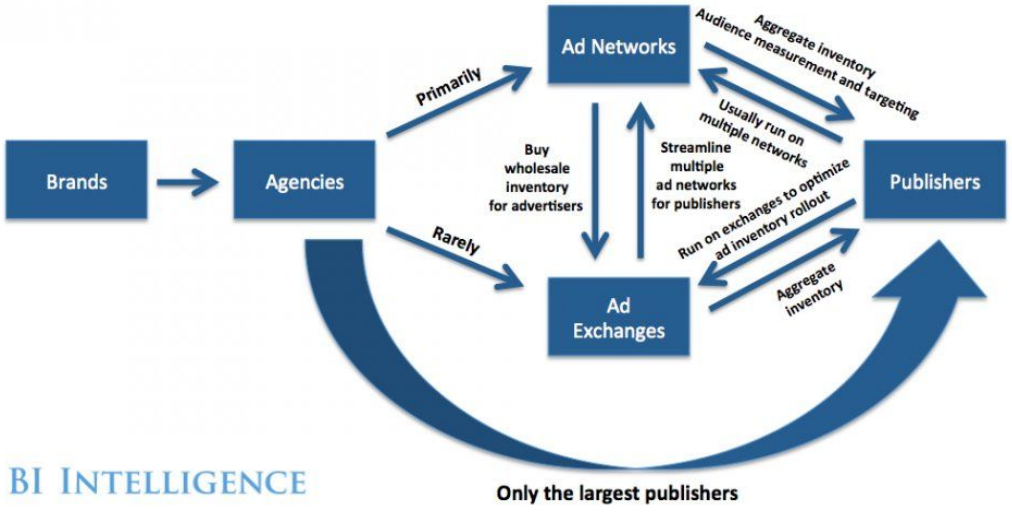
2.2. Redes de anúncios

Este modelo de monetização possui várias entidades envolvidas, e cada uma dessas entidades possui papéis e objetivos diferentes ao escolher monetizar com anúncios. Esse processo envolve, no geral, a rede de anúncio, os anunciantes, e o desenvolvedor que cede o espaço no seu aplicativo. De forma indireta, também envolve os usuários dos aplicativos, que são impactados com as propagandas, e interagem com as mesmas, mas eles não serão o foco desta seção.

As redes de anúncios servem como um intermediário entre os anunciantes e os desenvolvedores de aplicativos buscando monetização (*publishers*). Com o crescimento do digital, a publicidade passou a procurar espaços nesses meios para impactar usuários com suas mídias, e as redes de anúncios são responsáveis por facilitar essa veiculação de anúncios. Já do lado do desenvolvedor, com o crescimento da monetização por meio de anúncios, aumentou a procura por esse tipo de serviço, e os *publishers* passaram a contar com as redes de anúncios para poder preencher esses espaços cedidos nos seus aplicativos com mídia. Dessa maneira, as redes de anúncios conectam estas duas entidades, num modelo de negócio que gera lucros para estas redes também. As redes têm um funcionamento próprio, que será exposto mais adiante, com o objetivo de proporcionar um maior entendimento do cenário.

No diagrama abaixo, é possível visualizar como as redes de anúncios (*ad networks*) se posicionam entre os anunciantes ou agências de publicidade (*agencies*) e os desenvolvedores (*publishers*). É importante ressaltar que há outras entidades envolvidas no ecossistema de anúncios *mobile*, como as *Ad Exchanges,* mas elas não são o foco deste trabalho.

Figura 2: Ecossistema de anúncios mobile



Fonte: Business Insider [7]

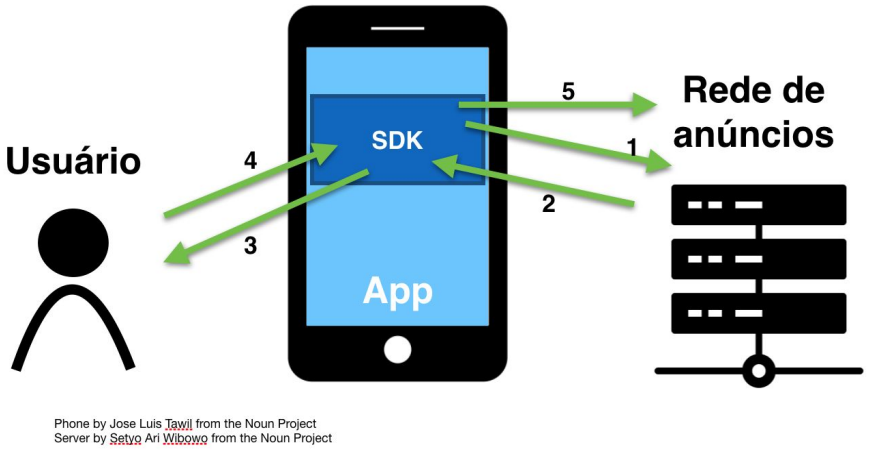
2.2.1. Funcionamento

Primeiramente, para haver a exibição de anúncios no aplicativo do desenvolvedor que deseja monetizar, é preciso que o aplicativo esteja conectado à rede de anúncios. No caso de aplicativos *Android* e iOS, isso é feito, geralmente, através da integração com o SDK da rede de anúncios utilizada. Um SDK [8] (*Software Development Kit*) é um conjunto de ferramentas e utilidades que auxiliam no desenvolvimento de uma aplicação, geralmente associada a algum ambiente específico. Nesse caso, o SDK da rede de anúncios disponibiliza ao desenvolvedor ferramentas capazes de exibir anúncios no seu aplicativo. O processo de integração também pode envolver cadastros nas redes de anúncios, e obtenção de *tokens*.

Uma vez que o *publisher* está devidamente integrado com a rede, ele precisa desenvolver no seu *app* a *feature* que será responsável por exibir esses anúncios. Esse processo geralmente envolve *designers* e testes com usuários, e abrange tanto a implementação, quanto a escolha do tipo de anúncio, local do anúncio na tela e frequência de exibição dos anúncios. Depois dessa implementação e do lançamento da nova versão do aplicativo na loja, os usuários passarão a ser impactados com as propagandas, e o desenvolvedor vai começar a monetizar.

O fluxo básico do funcionamento da requisição e exibição de anúncios envolvendo o SDK e a rede de anúncios está diagramado na imagem abaixo. Os passos estão numerados em ordem, e serão explicados mais adiante.

Figura 3: Fluxo básico de uma rede de anúncios



Fonte: Autora

Para exibir um anúncio, o SDK precisa se comunicar com a rede de anúncios, fazendo uma requisição de anúncio para a mesma, o *ad Request* (Representado no passo 1 da imagem). Na rede de anúncios, essa requisição é processada, e caso exista algum anúncio compatível com a requisição, o servidor devolve para o SDK um anúncio preenchido, ou seja, um *filled ad request.* Caso não exista um anúncio, o servidor não preenche a resposta, ou seja, envia um *no fill* para o SDK. Essa resposta da rede (*filled ad* ou *no fill*) é representada no passo 2 da imagem. Considerando que a resposta veio com um anúncio, o SDK exibe o anúncio na tela do aplicativo, e o usuário é impactado com a propaganda. Essa etapa é chamada de *impressão* (passo 3 da imagem). Dado que o anúncio está disponível para o usuário, o usuário pode clicar no mesmo (passo 4 da imagem), quando isso acontece, o SDK envia para a rede de anúncios uma requisição informando que houve um clique (passo 5 da imagem). Quando não há clique, ainda assim o SDK envia para a rede uma requisição informando que houve uma impressão.

É importante ressaltar que os SDK's e as redes de anúncios precisam ter uma arquitetura robusta, pois, como servem de agente intermediário entre publishers e anunciantes, é preciso ter um controle de todas as ações do SDK (requisições, impressões,

cliques) para fundamentar bem as métricas obtidas. São as métricas que irão definir o valor pago aos *publishers*, bem como o valor cobrado aos anunciantes, por isso é tão importante ter um bom controle desses eventos.

2.2.2. Métricas e monetização

Dado que já é conhecido o fluxo básico, é importante ressaltar alguns conceitos e informações relevantes para este trabalho. Como descrito acima, uma requisição para a rede de anúncios pode resultar em um *filled ad* ou em um *no fill.* A métrica denominada ***fill rate*** é a razão entre o número de *filled ads* e o número total de requisições. Então, quanto maior for o *fill rate* é melhor para o publisher, pois ele consegue fazer mais impressões de anúncios, e consequentemente obter um rendimento maior.

Quando o usuário clica num anúncio, é muito positivo para a marca que está veiculando a propaganda, pois significa que o usuário realmente foi impactado com o anúncio, pois demonstrou interesse ao clicar. Por isso, as redes costumam pagar mais ao *publisher* quando ocorre esse tipo de ação, então é de interesse do *publisher* que o maior número de usuários clique nos anúncios exibidos. As redes definem uma métrica para os cliques, que é o CTR, ou *Click Through Rate.* Essa métrica é dada pela razão entre o número de cliques e o número de impressões.

Outra métrica importante é o CPM, que corresponde ao ganho a cada mil impressões, ou seja, o ganho do publisher a cada mil impressões, que depende do CTR obtido. A fórmula para o CPM está abaixo, e a variável g representa o ganho do publisher a cada clique.

$$CPM = CTR \times g \times 10$$

Essa métrica é utilizada para medir o resultado de uma determinada campanha, de forma que quanto maior o CPM, maior foi o CTR, então maior foi impacto que a campanha teve, pois obteve muitos cliques. Quando a campanha tem um CPM alto, a rede de anúncios é beneficiada, pois ganha mais pela veiculação da mesma.

Dessa forma, é possível entender como as métricas influenciam nos lucros das redes de anúncios e dos desenvolvedores. Mais especificamente para os desenvolvedores, é possível perceber que é de interesse dos mesmos maximizar o número de impressões e o CTR.

2.2.3. Formatos dos anúncios

Dado que colocar anúncios em um aplicativo ocupa a tela, e pode influenciar na experiência do usuário, as redes de anúncios geralmente disponibilizam diversos formatos para os anúncios, com o objetivo de deixar o time de desenvolvimento do aplicativo mais livre para inserir estes espaços para propagandas dentro dos seus aplicativos. É imprescindível para uma boa experiência do usuário que o desenvolvedor seja cauteloso na escolha dos tipos e posicionamentos das propagandas, pois os anúncios podem vir a deixar o usuário insatisfeito ao ponto de deixar de usar o aplicativo.

Existem vários tipos de anúncios disponíveis, e um dos mais usados é o *banner*. O formato de *banner* é basicamente um espaço retangular, que possui tamanhos variados, e pode ser posicionado em qualquer espaço da tela. Geralmente o *banner* é utilizado nas extremidades da tela, ou seja, na parte superior ou inferior. Um fator a ressaltar é que, como o *banner* é pequeno em relação aos demais formatos, e geralmente está posicionado nas extremidades, o usuário tende a ignorá-lo, por isso esse é um formato que, no geral, rende menos cliques.

Outro formato muito utilizado é o anúncio de tela cheia, ou *interstitial.* Este tipo de anúncio, dependendo da implementação do desenvolvedor, pode surgir na tela mediante à alguma ação do usuário, ou em um momento qualquer. O anúncio ocupa a tela inteira, e possui um botão para fechar. Geralmente é usado em jogos, aparecendo após uma partida, por exemplo. Este tipo de anúncio, por ocupar a tela inteira e realmente ser visto pelo usuário, é um formato que possui melhores taxas de clique.

Um formato que vêm sendo bastante utilizado é o *native*. Este formato é customizável, e permite que o desenvolvedor altere suas propriedades (cores, fontes, botões, caixas de texto) e seu tamanho. Assim, o visual do anúncio fica parecido com o design do aplicativo, se tornando mais confortável e natural para o usuário. Além dos formatos citados, existem outros, como os anúncios com video e de notificação *push*.

3. Fraudes

De acordo com a Associação de Examinadores de Fraude Certificados (ACFE), a definição de fraude é [9]: o uso de sua ocupação para o enriquecimento pessoal através do uso indevido deliberado ou da aplicação dos recursos ou ativos da organização empregadora. Dada essa definição, fica claro que um suposto fraudador tem como objetivo principal obter algum benefício através de meios ilegais ou desonestos.

Atualmente as fraudes atingem os mais diversos setores. Mais especificamente no setor tecnológico, é muito comum encontrar casos de fraudes em redes de telecomunicações, *e-commerce*, bancos online, comunicações mobile, e em anúncios on-line, que é o foco deste trabalho. De acordo com os estudos de Kou Et. Al. [1], as fraudes estão aumentando dramaticamente com a expansão de tecnologia de ponta e da comunicação global, resultando em perdas substanciais para diversos negócios.

Como consequência disso, a detecção de fraudes tornou-se um tema de estudo bastante relevante, e vem sendo explorado na literatura. Este capítulo tem o objetivo de explorar as fraudes no contexto dos anúncios mobile e identificar técnicas para a detecção dessas fraudes.

3.1. Fraudes em anúncios mobile

Estima-se que em 2013 o mercado de anúncios *mobile* sofreu uma perda de quase um bilhão de dólares devido a comportamentos fraudulentos [2]. Possivelmente esse número aumentou nos anos seguintes. Como esse modelo de negócio compreende várias partes, como o *publisher*, o anunciante, a rede de anúncios, e o usuário final, sempre uma (ou mais) das partes é prejudicada quando há um comportamento fraudulento.

É importante salientar que existem diversos tipos de fraudes em anúncios *mobile*, e nem sempre elas acontecem por parte do *publisher*. Segundo um estudo realizado pela *Tune* [10], que envolveu o estudo de dados de mais de 700 redes de anúncios, algumas vezes a fraude parte da própria rede de anúncios. A análise mostrou que oito redes consistem apenas de cliques fraudulentos, e que 35 redes de anúncios possuem mais de 50% de fraudes nos

cliques. Porém, neste trabalho, o foco vai ser nos tipos de fraudes realizadas do lado do *publisher*, que são as mais comuns.

Depois de explorado o funcionamento de uma rede de anúncios, e como o desenvolvedor consegue obter lucros usando esse modelo, é possível perceber que é de interesse de desenvolvedores maliciosos cometer algum tipo de fraude com o objetivo de aumentar a base de lucros. Pelo fato de as redes de anúncio pagarem o *publisher* dependendo da quantidade de impressão de anúncios, ou da quantidade de cliques, ou até de uma combinação desses dois fatores, os *publishers* maliciosos tendem a utilizar técnicas para fraudar essas duas métricas. Nesta seção, serão apresentadas algumas formas de fraude.

3.1.1. Fraudes por execução de Cliques e Impressões

Como apresentado no Capítulo 2, as métricas relativas a cliques e impressões estão diretamente relacionadas com a renda que o *publisher* vai receber da rede de anúncios. Dessa maneira, quando um publisher mal intencionado deseja cometer algum comportamento fraudulento, é comum que se recorra a fraudar impressões e cliques de anúncios dentro do seu aplicativo.

Com o crescimento da indústria de anúncios *mobile* cresceu também o número de fraudes, e as técnicas usadas para fraudar as métricas evoluíram. Surgiu o conceito apelidado de *click farm,* "Fazenda de cliques" em tradução livre. Essas fazendas consistem em centenas ou milhares de dispositivos conectados a um computador, que são programados para fazer cliques no aplicativo 24h por dia. Esses cliques, no caso de fraudes em anúncios, têm o objetivo de provocar impressões de anúncios, e posteriormente clicar nos mesmos, gerando receita para o *publisher* do aplicativo em questão. Essas fazendas controladas por computadores também são conhecidas por "*Bot Farm*s", e além de cliques em anúncios, são responsáveis também por outras fraudes da indústria *mobile*, como por exemplo curtidas e seguidores em redes sociais, instalação de aplicativos, avaliação de aplicativos nas lojas, entre outros.

Segundo o estudo realizado por Cristofaro et. al. [11], após a análise de algumas fazendas concluiu-se que algumas dessas fazendas operadas por *bots* não tentam esconder a natureza das suas operações, enquanto outras, mais cuidadosas, seguem uma abordagem mais furtiva, tentando imitar o comportamento dos usuários regulares. Por esse motivo, é muito

difícil chegar a um consenso do que é fraude e do que não é, pois a indústria de fraudes está sempre evoluindo para simular o comportamento humano.

É importante salientar que essa prática é considerada criminosa, mas ainda ocorre com frequência. A maioria das *bot farms* estão localizadas na Tailândia (como a da figura a seguir), China e Rússia. Em dezembro de 2016 foi descoberto um esquema de fraudes em visualizações de vídeos na Rússia, no qual eram simulados 300 Milhões de visualizações, e onde os hackers lucraram cerca de 5 Milhões de dólares diariamente. De acordo com a Forbes [12], essa foi considerada a maior fraude de anúncios até então.

Figura 4: Fazenda de Cliques encontrada na Tailândia em Junho de 2017

Fonte: *Motherboard* [13]

Além das fazendas de clique controladas por computadores, existe também a prática de pagar pessoas para realizar esses cliques manualmente. É uma prática que custa mais caro para o fraudador, e que tem um alcance menor do que as fazendas de cliques, por ser manual. Mas, em contrapartida, é mais difícil de detectar que existiu um comportamento fraudulento justamente por se tratar de um ser humano fazendo este trabalho.

Outra técnica que pode ser usada para fraudar cliques e impressões é enviar requisições de um servidor diretamente para a rede de anúncios, como se fosse um usuário do aplicativo fazendo essa requisição via SDK. Essa técnica é mais rara, porque é mais difícil de ser executada, dado que é necessário um ataque do tipo *Man in the Middle*. Esse cenário

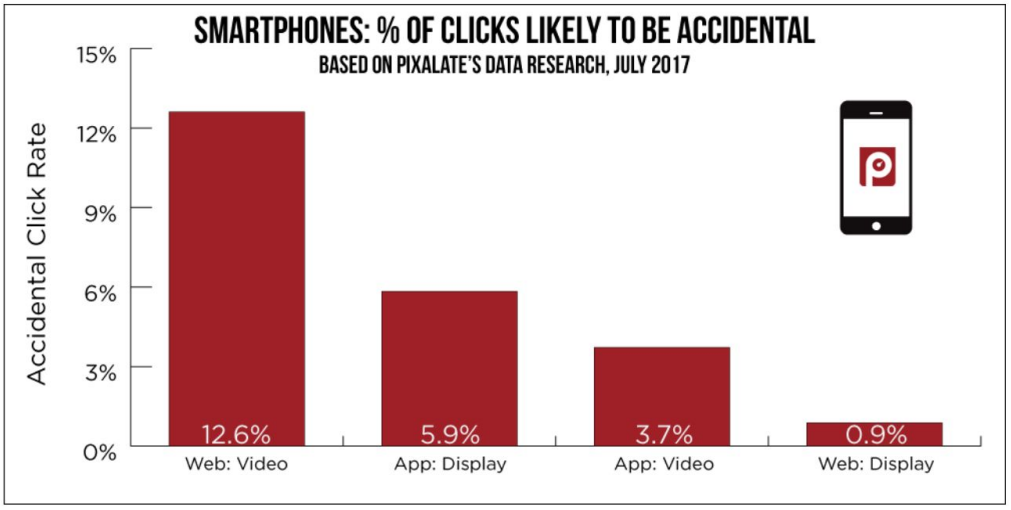
consiste em interceptar as requisições, para posteriormente entendê-las e simulá-las, ou ainda, alterar os dados do pacote enviado. Além disso, este tipo de ataque também depende da existência de vulnerabilidades no servidor da rede de anúncios, que precisa estar despreparado para lidar com requisições vindas de fora do SDK, e aceitar esse tipo de conduta.

3.1.2. Fraudes por posicionamento

Além das fraudes ocasionadas por fazendas de cliques, pessoas pagas para realizar cliques, e simulação de requisições, existe outra abordagem utilizada para fraudar anúncios, que se baseia no posicionamento do anúncio na tela do celular.

De acordo com o estudo de Liu Et Al. [3], que teve como foco fraudes desse tipo, fraudes de posicionamento manipulam *layouts* visuais de anúncios para desencadear impressões de anúncios e cliques involuntários de usuários reais. Esse tipo de fraude se baseia no fato de que o usuário pode fazer cliques acidentais, e o fraudador tenta fazer com que isso aconteça. Um estudo da empresa *Pixalate* [14] apontou que quase 6% dos cliques em anúncios em aplicativos *mobile* são acidentais. Na figura abaixo, estão dispostas algumas estatísticas desse estudo.

Figura 5: Índices dos cliques Acidentais em Anúncios



Fonte: *Pixalate* [14]

Esse tipo de fraude pode ser realizado de diversas maneiras: colocando mais de um anúncio na mesma posição, sobrepondo os demais; esconder o anúncio atrás de outros elementos da tela, como botões, textos e imagens; re-dimensionamento de anúncios para dificultar a visibilidade dos mesmos. Todas essas técnicas têm o objetivo de aumentar o número de impressões e de ocasionar cliques acidentais.

Vale ressaltar que a maioria das redes de anúncios usam a métricas de *viewability* para contabilizar impressões. Essas métricas são relacionadas à percentagem do *frame* do anúncio que está sendo realmente exibida para o usuário, e também à quantidade de tempo desde o carregamento do anúncio. Dessa forma, esse tipo de fraude fica mais restrito em redes que possuem políticas de *viewability*.

Dado que as fraudes são um problema existente no cenário em questão, é importante que as redes tenham conhecimento sobre esse tipo de comportamento por parte dos *publishers*, por isso é importante detectar esses casos. Na seção a seguir será apresentado o estado da arte sobre detecção de fraudes.

3.2. Detecção de fraudes

Na literatura existem diversos autores que estudam a detecção de fraudes. Por isso, já foram propostas diversas técnicas envolvendo conceitos de várias áreas da computação e da estatística. É importante salientar que a escolha da técnica mais apropriada varia caso a caso, de acordo com o tipo de fraude a ser detectada, e também com o teor dos dados a serem analisados.

Segundo Kou Et. Al. [1], em relação a fraudes em cartão de crédito, métodos não supervisionados não requerem um conhecimento prévio das transações fraudulentas e não fraudulentas da base de dados, mas, pelo contrário, detectam mudanças de comportamento ou transações não usuais. Esses métodos modelam uma distribuição de *baseline* que representa o comportamento normal e, em seguida, detecta as observações que mostram a maior saída dessa norma.

Apesar de o foco do trabalho não ser cartões de crédito, os dados relativos aos anúncios (que são basicamente relativos à requisições, cliques e impressões) tem um viés muito parecido com os citados por Kou, pois não há conhecimento sobre as requisições,

cliques e impressões fraudulentas. Dessa maneira, foram escolhidas técnicas parecidas com a descrita acima para o desenvolvimento deste trabalho.

3.2.1. Detecção de *outliers*

De acordo com a definição de Hawkins [15], em 1980, um *outlier* pode ser definido como uma observação que se desvia tanto das outras observações, ao ponto de levantar suspeitas de que foi gerada por um mecanismo diferente. Ou seja, em um conjunto de dados, um *outlier* vai ser um elemento que é muito diferente dos demais, ou que não segue o mesmo padrão observado.

O termo "mecanismo diferente" pode englobar diversas explicações. Para Hodge [16], *outliers* surgem devido a erros humanos, erros de instrumentos, desvios naturais em populações, comportamento fraudulento, mudanças no comportamento de sistemas ou falhas em sistemas.

Dessa forma, podemos observar que, de forma geral, os comportamentos fraudulentos acabam gerando barulho nos dados, que podem ser identificados como *outliers* depois de uma análise acerca da distribuição dos mesmos.

Existem diversas técnicas desenvolvidas para a detecção de *outliers*, e elas variam de acordo com o conhecimento que existe sobre os dados anômalos da base. A escolha da técnica também depende das variáveis do conjunto de dados e da relação entre elas, a análise pode ser pautada em dados univariados ou multivariados. Em relação aos algoritmos, a maioria das técnicas utilizadas têm um embasamento estatístico, baseando-se na distribuição dos dados, na distância entre eles quando dispostos em algum tipo de gráfico, entre outros. Os métodos desenvolvidos mais recentemente fazem uso de redes neurais e de aprendizagem de máquina para a detecção de *outliers*, mas, no geral, essas técnicas dependem de um conjunto de dados rotulado, ou que pelo menos uma parte deles esteja rotulada.

3.2.2. O uso da detecção de *outliers* na identificação de fraudes

As áreas de detecção de *outliers* e detecção de fraudes estão muito ligadas, pois a maioria dos estudos sobre detecção de fraudes sugere a detecção de *outliers* para a

identificação de fraudes, justamente pelo fato de uma fraude, em geral, ser um ponto fora da curva num determinado conjunto de dados.

De acordo com Hodge [16], numa base de dados, *outliers* podem indicar casos fraudulentos, ou eles podem apenas denotar um erro, ou uma interpretação errada de um valor em falta, de qualquer forma, a detecção da anomalia é vital para a consistência e a integridade da base de dados. Dessa maneira, é importantíssimo detectar esses *outliers*, pois eles podem vir a indicar uma fraude.

A recíproca também é verdadeira, pois em muitos estudos sobre detecção de fraudes a maioria das técnicas envolve apontar *outliers* num conjunto de dados. Os estudo realizado por Kou et. Al. [1] analisou diferentes tipos de fraudes, e citou os métodos de detecção mais apropriados para cada uma delas. Os autores ressaltaram que a técnica depende do conhecimento dos dados, ou seja, se, em um conjunto de dados, é conhecido que uma observação que é fraudulenta. Os autores frisaram que para um cenário onde não existe conhecimento sobre a observação é recomendado o uso de técnicas de detecção de *outliers* que se baseiam na distribuição dos dados ou na distância de uma determinada observação para as demais.

Dado que fraudes são observações que desviam do comum, é possível entender que possivelmente os dados relativos a fraudes na base de dados provavelmente são *outliers*, quando analisados em meio ao conjunto inteiro. Desse modo, é válido dizer que um estudo dos *outliers* de um certo conjunto de dados pode trazer conhecimento acerca da ocorrência de fraudes num determinado sistema. Neste trabalho será feita uma análise dos dados de impressões e cliques de uma rede de anúncios, com o objetivo de traçar os *outliers*, e indicar possíveis fraudes no contexto estudado.

4. Desenvolvimento

Este capítulo compreende todo o desenvolvimento do trabalho, desde a obtenção dos dados que foram analisados, até os resultados e as análises comparativas. Primeiramente, na seção 4.1, é descrito todo o processo de obtenção e manipulação dos dados, e são apresentados os conjuntos de dados utilizados no desenvolvimento. Nas seções seguintes são feitas as análises dos dados em busca da identificação de *outliers*.

Uma vez que os dados não são classificados, é recomendado o uso de técnicas que envolvam analisar os valores comparando-os com uma distribuição e, em seguida, detectar as observações que mostram a maior saída dessa norma [1]. Por isso foram escolhidas técnicas com esta natureza para fazer a análise. Na seção 4.2 é feita uma análise baseada na comparação de distribuições estatísticas, já na seção 4.3 é disposta uma análise estatística baseada em valores espúrios.

A maioria das técnicas utilizadas na literatura são adequadas para distribuições que seguem uma distribuição estatística normal. Como neste trabalho nenhuma das variáveis segue esse tipo de distribuição, não foi possível utilizar alguns métodos clássicos como o Z-Score, por exemplo.

4.1. Obtenção dos dados

Os dados utilizados neste trabalho são dados reais de uma rede de anúncios *mobile*. Como visto no capítulo 2, redes de anúncios funcionam se comunicando com os dispositivos do usuário através do SDK da rede de anúncios, e no fluxo desde a requisição do anúncio até o clique no anúncio acontecem várias ações, que são registradas. Todo esse fluxo gera informações relativas a requisições de anúncios, respostas, cliques, impressões, horários dos acontecimentos, informações do dispositivo que solicitou o anúncio, dados de localização, entre muitos outros. Por esse motivo, o conjunto de dados é gigantesco, principalmente se for considerado um longo período de tempo.

Todos esses dados coletados estão espalhados em algumas tabelas, e cada uma dessas tabelas têm diversas variáveis a serem analisadas. É importante ressaltar que no conjunto de

dados em questão não há conhecimento relativo a usuários ou aplicativos fraudadores, e nem a informação de se uma dada requisição ou ação foi caracterizada ou não uma fraude. Um ponto positivo é que, por serem dados reais, provavelmente existem observações que foram geradas por comportamento fraudulento, mas que não há conhecimento ainda. Um dos objetivos dos experimentos realizados é justamente indicar que determinado usuário ou aplicativo é um candidato para ser considerado um fraudador.

Num primeiro momento foram coletados dados brutos relativos às requisições de anúncios. Nesse conjunto de dados existiam diversas variáveis, como: identificador do dispositivo do usuário, identificador do aplicativo, horário, sistema operacional, modelo do dispositivo, entre muitas outras. Cada linha desse conjunto representava uma requisição de anúncio, por isso, mesmo com muitas linhas, como existiam muitas requisições de muitos usuários, quando se separava esses dados por usuário, não havia muita informação relevante. Após alguns estudos sobre esse conjunto de dados em particular, e de se observar as relações entre as variáveis, foi decidido mudar a abordagem.

Após a coleta inicial, houve uma segunda fase de coletas: desta vez o objetivo foi agregar os dados de uma forma que fosse possível extrair métricas relevantes para o estudo. Dessa maneira, recorreu-se a um conjunto de dados que possui informações da requisição relativas a ações realizadas: clique ou impressão. Nesta segunda etapa os dados foram coletados já com agregações por usuários, ou seja, o conjunto de dados obtido consistia em diversas linhas, cada uma com um usuário e os valores de cada variável para este usuário. Dessa forma, foi possível ter a contagem de cliques e de impressões para cada usuário em um determinado intervalo de tempo. A agregação foi importante pois cliques e impressões são justamente as variáveis mais relevantes para a análise, e ter essa informação já contada e agregada foi um facilitador para o desenvolvimento. Foi escolhido o intervalo de tempo de três meses para a análise, pois era necessário um intervalo de tempo maior, dado que a análise é baseada em um comportamento.

Depois do início realização dos experimentos, foi constatada a necessidade de alguns filtros para os dados coletados. O intervalo de tempo de três meses para os dados coletados gerou dados um pouco enviesados, pois no mesmo conjunto havia usuários que estavam presentes desde o primeiro dia, e usuários que só apareceram no último dia. Por esse motivo foi gerado um novo conjunto de dados, mas dessa vez, apenas com usuários que realizaram alguma ação em pelo menos 45 dos 90 dias analisados. Uma vez que essa filtragem foi feita,

os novos dados se mostraram mais homogêneos do que os anteriores, pois este novo conjunto compreendeu usuários comprovadamente ativos.

Ao longo dos experimentos também foi identificada a necessidade de outros conjuntos de dados. As variáveis analisadas continuaram sendo as mesmas (número de cliques e impressões), mas agora foi gerado um conjunto de dados com o número de cliques e impressões de cada aplicativo no período de três meses. Além desse novo conjunto agrupado por aplicativo, foram gerados 5 novos conjuntos agrupados por usuário, cada um deles relativo a um aplicativo diferente, sendo os aplicativos escolhidos aqueles com um maior volume de requisições. Os nomes dos aplicativos não podem ser divulgados devido a questões de privacidade dos dados da empresa, para fins de compreensão, durante o trabalho os aplicativos serão referenciados por um número.

Os conjuntos de dados utilizados neste trabalho são descritos e nomeados a seguir. Todos os conjuntos correspondem a um intervalo de tempo de três meses e consideram apenas usuários ativos. Vale ressaltar que os nomes utilizados nesta seção serão usados ao longo do trabalho.

- **Conjunto de dados principal:** Este conjunto compreende 10 mil entradas. Cada uma delas possui as seguintes informações: identificador do usuário, número de cliques do usuário, número de impressões do usuário e taxa de clique (razão entre o número de cliques e o número de impressões). Este conjunto está separado por usuário, e compreende usuários de vários aplicativos.

- **Conjunto de dados por aplicativo:** Este conjunto compreende quase 560 entradas. Cada uma delas possui as seguintes informações: identificador do aplicativo, número de cliques desse aplicativo, número de impressões desse aplicativo e taxa de clique. Este conjunto está separado por aplicativo, e agrega todas as ações de um dado aplicativo em uma entrada.

- **Conjuntos de dados adicionais:** Os conjuntos adicionais são cinco no total, cada um deles com 10 mil entradas. A natureza dos dados de cada um destes conjuntos é igual à do conjunto de dados principal descrito mais acima, a diferença é que para cada um dos cinco conjuntos só existem dados de um aplicativo. Ou seja, cada conjunto

representa dados exclusivos de um aplicativo, nesse caso, os cinco aplicativos com mais requisições.

As tecnologias usadas nesta etapa de coleta de dados foram basicamente uma API que acessa o banco de dados, e *queries* SQL que possibilitaram os agrupamentos citados.

4.2. Análise das distribuições

O passo seguinte no desenvolvimento foi analisar a distribuição dos dados coletados com o objetivo de identificar um possível aplicativo fraudador. De acordo com Hawkins [15], no seu estudo sobre identificação de *outliers*, os métodos paramétricos estatísticos para a identificação de outliers assumem uma distribuição conhecida das observações. Por esse motivo foi feita uma análise acerca da distribuição dos dados coletados, com o objetivo de identificar algum padrão na distribuição dos dados. É importante ressaltar que esse conjunto de dados provavelmente possui observações que são outliers, por se tratar de um conjunto de dados real. Por isso, o conjunto possui ruídos, o que pode prejudicar a análise.

Após a identificação da distribuição dos dados para cada uma das três variáveis estudadas (taxa de clique, número de cliques e número de impressões), por meio de testes com o conjunto de dados principal, é possível fazer uma comparação com as distribuições de cada conjunto adicional (dados referentes a apenas um aplicativo). Após a comparação, será possível identificar anomalias e apontar possíveis aplicativos fraudadores.

4.2.1. Identificação das distribuições

Dado que o objetivo é fazer uma análise comparativa entre as distribuições, é preciso obter o ajuste da distribuição para cada uma das variáveis. O ajuste de distribuição é uma prática comum na estatística, e consiste em encontrar o tipo da distribuição (Normal, Exponencial, Lognormal, etc.) e os valores dos parâmetros (média, desvio padrão, etc.) que dão a maior probabilidade de produzir os dados observados [17].

É importante ressaltar que os valores utilizados para esta identificação são os do conjunto de dados principal, ou seja, são dados agrupados por usuário, e compreendendo usuários de diversos aplicativos. O objetivo desta etapa é traçar os valores que serão utilizados como base nas etapas seguintes, e são valores válidos por compreenderem uma gama grande de aplicativos. Uma vez que esse dado é obtido, a análise comparativa das distribuições das variáveis por aplicativo será feita sempre comparando esses valores base com os valores da distribuição correspondente ao aplicativo em questão.

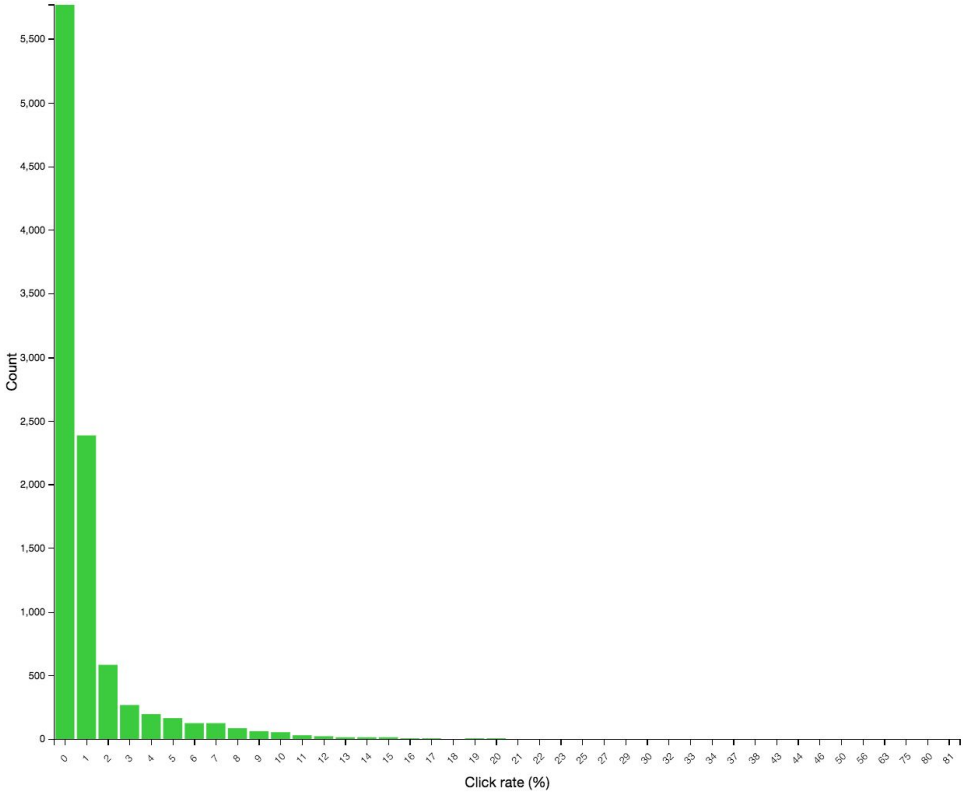
Para cada variável a ser analisada, a metodologia foi: plotar um gráfico com a distribuição dos dados, criar hipóteses sobre o tipo de distribuição que os dados seguem, aplicar métodos estatísticos de ajuste da distribuição (*Maximum-likelihood Fitting*) para tentar comprovar a hipótese e obter o tipo de distribuição mais adequado e os valores dos seus parâmetros.

4.2.1.1. Taxa de Clique

A variável de taxa de clique, que é a razão entre o número de cliques e o número de impressões. Dentre as três variáveis estudadas, essa é a mais importante, pois é a que tem um maior valor semântico entre as três (taxa de clique, número de cliques e número de impressões). Por este motivo, essa variável vai ser o principal indicador de fraudes ao analisar os resultados.

A análise inicial desta variável consistiu em plotar um gráfico de distribuição dos dados. Para esta etapa, as tecnologias utilizadas foram Javascript, e duas bibliotecas voltadas para a visualização de dados: DC.js, que facilita a construção de gráficos, e Crossfilter.js, que simplifica a manipulação, agrupamento, e filtragem dos dados. O gráfico correspondente à distribuição da taxa de clique é a figura abaixo.

Figura 6: Distribuição dos valores da Taxa de Cliques do Conjunto Principal



Fonte: Autora

Ao observar o gráfico é possível identificar que a distribuição se parece com uma curva do tipo exponencial. Porém, não é razoável assumir que a variável possui uma distribuição exponencial, é preciso fazer testes estatísticos e comprovar essa hipótese. Por isso, foram realizados ajustes de máxima probabilidade (*Maximum-likelihood Fitting*) para atingir a distribuição correta e os valores dos seus parâmetros.

Existe um *software* denominado *Free Statistics and Forecasting Software* [18] que é voltado para análises estatísticas, incluindo ajustes de máxima probabilidade para diversos tipos de distribuição. Para obter o ajuste de uma determinada distribuição, basta escolher a distribuição desejada e fornecer os dados. O *software* faz um processamento (que utiliza a linguagem R) e gera um relatório com os parâmetros para a distribuição em questão, ou um erro caso não seja possível ajustar os valores observados na distribuição em questão.

Para esta variável foram testadas algumas distribuições: exponencial, lognormal, e *gamma*. A distribuição que obteve um melhor resultado foi a exponencial, com o ajuste de máxima probabilidade exponencial. Foi concluído que os dados seguem uma distribuição

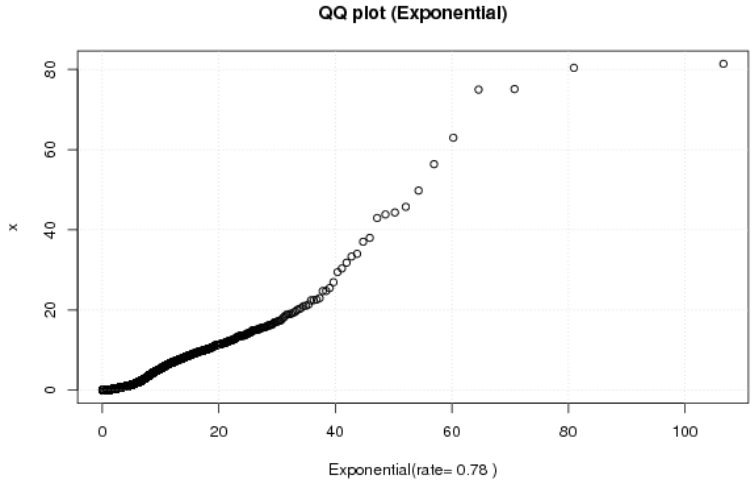
exponencial cuja taxa (λ) é de aproximadamente 0.78. Os resultados obtidos estão dispostos na tabela a seguir.

Tabela 1: Resultado do ajuste de distribuição exponencial para a Taxa de Cliques do Conjunto Principal

Parâmetro	**Valor Estimado**	**Desvio Padrão**
Taxa	0.778593182390314	0.00778593182390314

O software também gera um QQ plot, que é um gráfico útil para comparar duas distribuições de probabilidade, plotando os seus quantis nos eixos. Quanto mais próximos da reta $x = y$ os pontos estiverem, mais os dados se assemelham ao tipo de distribuição em questão. Esse gráfico está disposto abaixo.

Figura 7: QQ plot do ajuste exponencial para a Taxa de Cliques do Conjunto Principal



Fonte: *Free Statistics and Forecasting Software* [18]

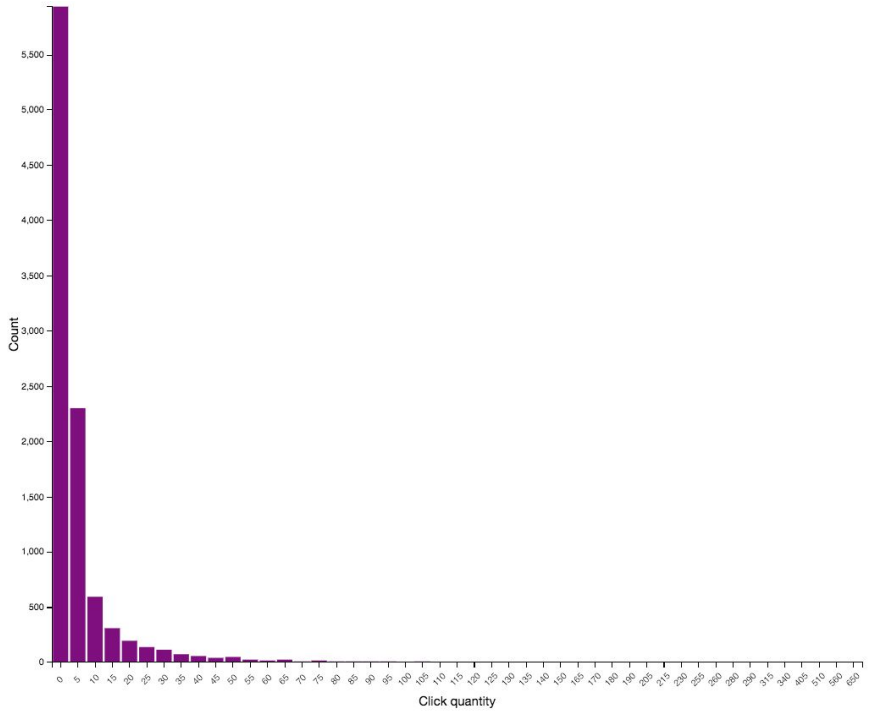
O valor da taxa, que é o parâmetro da distribuição exponencial (o seu *lambda*), será utilizado posteriormente, pois servirá como base para a comparação das distribuições específicas por aplicativo com esta distribuição geral.

4.2.1.2. Número de cliques

A segunda variável analisada é a do número de cliques, ou seja, em quantos anúncios um determinado usuário clicou. Semanticamente essa variável é menos significativa para a análise, pois esse valor é muito relativo, e depende de o quanto a pessoa utiliza o *app*. Muitos cliques podem significar uma fraude, mas dependendo do número de impressões para a mesma pessoa, um número alto de cliques pode ser normal. Por isso, é preciso de cuidado ao fazer inferências ao se basear nesta variável, então ela terá um peso menor nas análises posteriores.

Foi feita uma abordagem semelhante à da variável anterior, e foi inicialmente plotado o gráfico com a distribuição dos dados. O gráfico encontra-se na figura a seguir.

Figura 8: Distribuição dos valores do número de Cliques do Conjunto Principal



Fonte: Autora

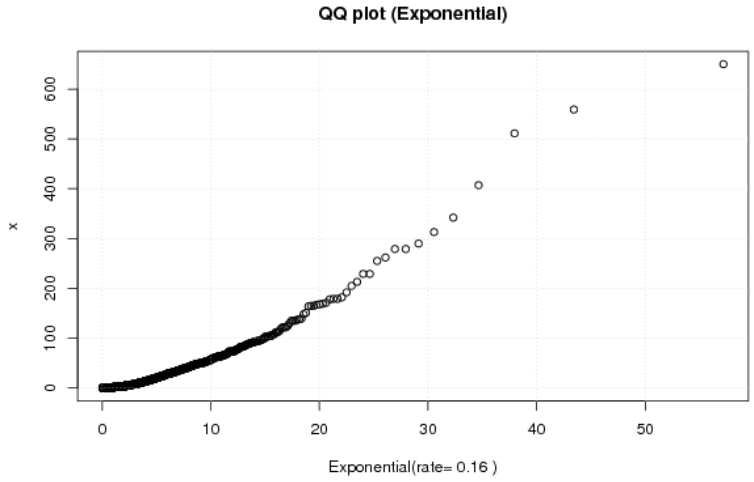
Após a análise do gráfico é possível levantar a hipótese de que os dados desta variável seguem uma distribuição do tipo exponencial. Após levantar essa hipótese, foram realizados alguns testes estatísticos de ajuste de distribuição (*Maximum-likelihood Fitting*), utilizando o mesmo *software* citado na subseção anterior. Foram feitos ajustes para diversas distribuições: exponencial, *gamma* e lognormal. Após os testes, a distribuição que se encaixou melhor foi realmente a exponencial. Os resultados estão dispostos a seguir.

Tabela 2: Resultados do ajuste de distribuição exponencial para o número de Cliques do Conjunto Principal

Parâmetro	**Valor Estimado**	**Desvio Padrão**
Taxa	0.155840917591323	0.00155840917591323

Pode-se perceber que no caso desta variável o valor da taxa para a distribuição exponencial foi de aproximadamente 0.16. Esse valor do λ é baixo comparado ao valor da variável de taxa de cliques (0.78), porém, das distribuições testadas, a exponencial foi a que se encaixou melhor para o número de cliques. O gráfico do QQ plot correspondente encontra-se abaixo.

Figura 9: QQ plot do ajuste exponencial para o Número de Cliques do Conjunto Principal



Fonte: *Free Statistics and Forecasting Software* [18]

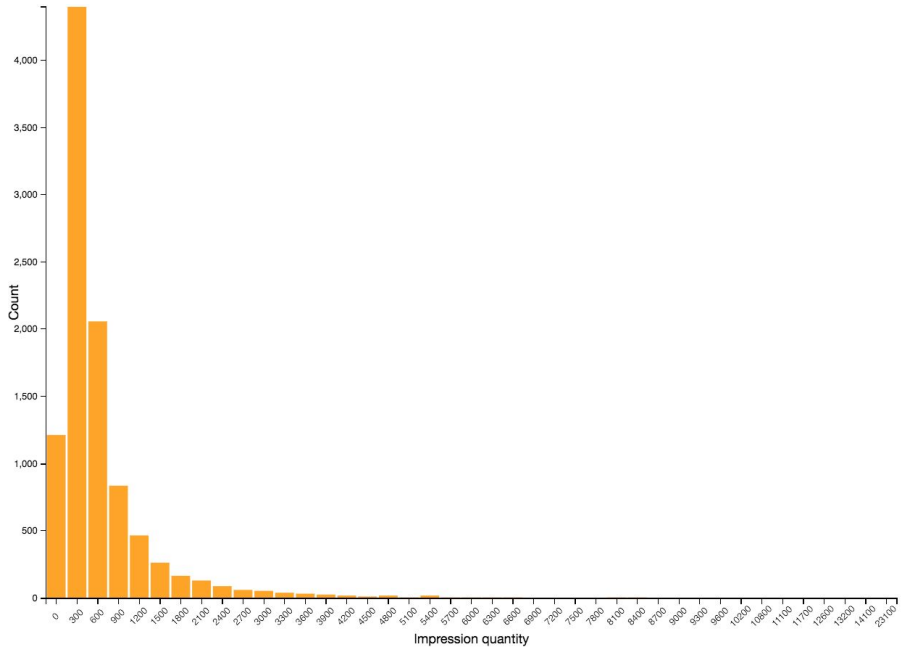
O valor da taxa, que é o parâmetro da distribuição exponencial (o seu *lambda*) será utilizado posteriormente, pois servirá como base para a comparação das distribuições específicas por aplicativo com esta distribuição geral.

4.2.1.3. Número de impressões

A última variável do estudo é o número de impressões, ou seja, quantos anúncios foram impressos para o determinado usuário no período de tempo analisado. Assim como o número de cliques, o valor semântico desta variável não é tão grande quanto o da taxa de cliques, pois é muito relativo, e depende da frequência de uso do usuário. Por esse motivo, ela também tem um peso menor na análise final.

Assim como com as demais variáveis, foi inicialmente plotado o gráfico com a distribuição dos dados. O gráfico encontra-se na figura a seguir.

Figura 10: Distribuição dos valores do Número de Impressões do Conjunto Principal



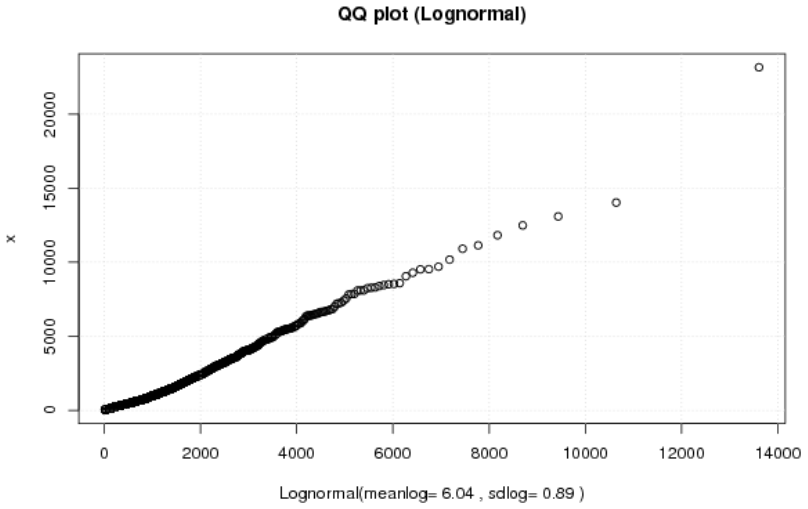
Fonte: Autora

O gráfico dessa variável diferiu dos das outras variáveis, pois, inicialmente, não se parece com uma distribuição exponencial. Com esta variável não foi feita uma hipótese sobre a natureza da sua distribuição, após examinar o gráfico, foram realizados os experimentos com os ajustes estatísticos para os diferentes tipos de distribuição. Os experimentos (*Maximum-likelihood Fitting*) foram realizados utilizando o mesmo *software*, e foram testadas a distribuição exponencial, lognormal, *gamma* e normal. A distribuição que demonstrou o melhor resultado para esta variável foi a lognormal. Os resultados e o QQ plot encontram-se a seguir.

Tabela 3: Resultados do ajuste de distribuição lognormal para o Número de Impressões do Conjunto Principal

Parâmetro	**Valor Estimado**	**Desvio Padrão**
Média do logaritmo	6.04263895076141	0.00893391815515167
Desvio padrão do logaritmo	0.893391815515167	0.00631723411007336

Figura 11: QQ plot do ajuste exponencial para o Número de Impressões do Conjunto Principal



Fonte: *Free Statistics and Forecasting Software* [18]

4.2.2. Análise comparativa

Após a obtenção da distribuição estatística e dos seus respectivos parâmetros para cada variável, foi possível traçar valores base para os dados analisados. Esses valores base serão considerados o padrão geral dos dados, por terem sido calculados usando o conjunto de dados principal, que contém usuários de diversos aplicativos.

Contando com esses valores base é possível fazer uma análise comparativa para cada um dos cinco aplicativos do conjunto de dados adicional. Cada um dos conjuntos conta com dados distribuídos por usuário, mas contam apenas com usuários do aplicativo em questão, ou seja, é um conjunto composto apenas por usuários do aplicativo cuja distribuição está sendo analisada. O que será feito nesta etapa é a comparação das distribuições das variáveis de cada um dos aplicativos com a distribuição base, e será feita uma análise dos resultados obtidos. Para a comparação das distribuições será utilizada a divergência de Kullback-Leibler.

4.2.2.1. Valores das distribuições para cada aplicativo

Para ser possível a comparação, é preciso fazer o ajuste de distribuição com cada variável de cada aplicativo analisado. Como já foi feita a análise da distribuição de cada variável com o conjunto base, já é conhecida a distribuição para cada variável. Então, nesta etapa foram realizados os ajustes para cada variável (taxa de cliques, número de cliques, e número de impressões), sendo utilizados para a taxa de cliques e o número de cliques um ajuste de máxima probabilidade exponencial, e para o número de impressões um ajuste de máxima probabilidade lognormal.

Foram realizados os testes para cada variável de cada aplicativo estudado, e os resultados encontram-se a seguir.

- **Taxa de Cliques**

Resultados para a taxa de cliques, realizando um ajuste de máxima probabilidade **exponencial**:

Tabela 4: Resultado do ajuste de distribuição exponencial para a Taxa de Cliques de cada aplicativo analisado

Aplicativo	Parâmetro	Valor Estimado	Desvio Padrão
Aplicativo 1	**Taxa**	1.32159503204371	0.0132159503204371
Aplicativo 2	**Taxa**	1.93187053267224	0.0193187053267224
Aplicativo 3	**Taxa**	1.18532182338487	0.0118532182338487
Aplicativo 4	**Taxa**	3.35131576876541	0.0335131576876541
Aplicativo 5	**Taxa**	0.959248292207864	0.00959248292207864

- **Número de Cliques**

Resultados para o número de cliques, realizando um ajuste de máxima probabilidade **exponencial**:

Tabela 5: Resultado do ajuste de distribuição exponencial para o Número de Cliques de cada aplicativo analisado

Aplicativo	Parâmetro	Valor Estimado	Desvio Padrão
Aplicativo 1	**Taxa**	0.632671137542705	0.00632671137542705
Aplicativo 2	**Taxa**	0.48661800486618	0.0048661800486618
Aplicativo 3	**Taxa**	0.503829101168883	0.00503829101168883
Aplicativo 4	**Taxa**	0.765872711955273	0.00765872711955273
Aplicativo 5	**Taxa**	1.09481059776659	0.0109481059776659

- **Número de Impressões**

Resultados para o número de impressões, realizando um ajuste de máxima probabilidade **lognormal**:

Tabela 6: Resultados do ajuste de distribuição lognormal do Número de Impressões de cada aplicativo analisado

Aplicativo	Parâmetro	Valor Estimado	Desvio Padrão
Aplicativo 1	**Média do logaritmo**	5.26746562851599	0.00838752070812949
	Desvio padrão do logaritmo	0.838752070812949	0.00593087277006095

Aplicativo 2	**Média do logaritmo**	5.86126120843588	0.00823829898633295
	Desvio padrão do logaritmo	0.823829898633295	0.00582535707867829
Aplicativo 3	**Média do logaritmo**	5.43110599950375	0.0053169403025659
	Desvio padrão do logaritmo	0.53169403025659	0.0037596445431084
Aplicativo 4	**Média do logaritmo**	6.07851199501852	0.00664954482681339
	Desvio padrão do logaritmo	0.664954482681339	0.00470193823884368
Aplicativo 5	**Média do logaritmo**	4.63501917158636	0.00712410730687765
	Desvio padrão do logaritmo	0.712410730687765	0.00503750458659382

4.2.2.2. Divergência de Kullback-Leibler

Ao realizar uma análise comparativa, além dos valores, é preciso de uma técnica eficiente para obter os resultados da comparação. Para este trabalho foi escolhido o cálculo da divergência de Kullback-Leibler. Esta divergência foi definida por Kullback e Leibler em 1951 [19], e é uma medida baseada na informação da disparidade entre as distribuições de probabilidade [20].

Também chamada de entropia relativa, essa medida é muito utilizada na estatística para comparar duas distribuições de probabilidade. Ao calcular o valor da divergência, quanto mais o valor se aproxima de zero, mais é possível esperar um comportamento semelhante para as duas distribuições. Caso o valor se aproxime de um, em contrapartida, é um indicativo de que as duas distribuições comparadas se comportam de uma forma diferente.

Para cada tipo de distribuição de probabilidade a divergência de Kullback-Leibler é calculada de uma maneira distinta, pois cada tipo de distribuição possui os seus próprios parâmetros. Vale ressaltar que dependendo da distribuição analisada e dos valores das mesmas os intervalos em que os valores estão compreendidos podem variar, porém, um valor próximo de zero sempre vai indicar similaridade, enquanto valores próximos ou acima de um vão indicar discrepância. Neste trabalho, serão utilizadas duas fórmulas, pois serão computadas divergências para distribuições exponenciais e distribuições lognormais.

A fórmula da divergência de Kullback-Leibler para uma distribuição Exponencial [21] é:

$$D(\lambda_0 \parallel \lambda) = log(\lambda_0) - log(\lambda) + \frac{\lambda}{\lambda_0} - 1$$

Na fórmula acima, a variável λ_0 representa a taxa da distribuição exponencial base. A variável λ é a taxa da distribuição que está sendo comparada com a distribuição base.

A fórmula da divergência de Kullback-Leibler para uma distribuição Lognormal [22] é:

$$D(f_i \parallel f_j) = \frac{1}{2\sigma_j^2}\left[(\mu_i - \mu_j)^2 + \sigma_i^2 - \sigma_j^2\right] + ln\frac{\sigma_j}{\sigma_i}$$

Para esta fórmula, i é considerada a distribuição lognormal base, e j é considerada a distribuição que está se comparando com a base. A variável σ representa o desvio padrão do logaritmo da distribuição, enquanto μ representa a média do logaritmo da distribuição.

4.2.2.3. Resultados

Ao aplicar a divergência de Kullback-Leibler para cada variável de cada aplicativo analisado, foram obtidos os resultados. Comparou-se, para cada variável de cada aplicativo, os valores dos parâmetros da distribuição correspondente (encontrados nas tabelas da subseção 4.2.2.1) e os valores base (da seção 4.2.1). As tabelas com os resultados obtidos para cada aplicativo encontram-se a seguir.

Tabela 7: Resultados do cálculo das divergências das distribuições para cada variável do aplicativo 1

Aplicativo 1	
Variável	**Divergência de Kullback-Leibler**
Taxa de Clique	0.4653488104347807
Número de Cliques	2.592720820276921

Número de impressões	0.18178638958152735

Tabela 8: Resultados do cálculo das divergências das distribuições para cada variável do aplicativo 2

Aplicativo 2	
Variável	**Divergência de Kullback-Leibler**
Taxa de Clique	1.0828736777298502
Número de Cliques	1.7330231177089694
Número de impressões	0.02793254863175839

Tabela 9: Resultados do cálculo das divergências das distribuições para cada variável do aplicativo 3

Aplicativo 3	
Variável	**Divergência de Kullback-Leibler**
Taxa de Clique	0.3379016855783554
Número de Cliques	1.8326686847790548
Número de impressões	0.3907442339247868

Tabela 10: Resultados do cálculo das divergências das distribuições para cada variável do aplicativo 4

Aplicativo 4	
Variável	**Divergência de Kullback-Leibler**
Taxa de Clique	2.663437930595598
Número de Cliques	3.397752742870317
Número de impressões	0.2138104387900272

Tabela 11: Resultados do cálculo das divergências das distribuições para cada variável do aplicativo 5

Aplicativo 5	
Variável	**Divergência de Kullback-Leibler**
Taxa de Clique	0.13996907084369603
Número de Cliques	5.435489584770596

Número de impressões	0.3505710082725811

Aplicativo	Variável	Divergência
Aplicativo 4	Taxa de Clique	2.663437930595598
	Número de Cliques	3.397752742870317
	Número de impressões	0.2138104387900272
Aplicativo 5	Taxa de Clique	0.13996907084369603
	Número de Cliques	5.435489584770596
	Número de impressões	0.3505710082725811
Aplicativo 3	Taxa de Clique	0.3379016855783554
	Número de Cliques	1.8326686847790548
	Número de impressões	0.3907442339247868

Ao fazer uma análise dos resultados, é possível perceber que, de forma geral, as divergências foram altas. A variável que obteve o melhor resultado, ou seja, obteve os resultados mais próximos de zero, que significa que a distribuição se comporta de forma semelhante à distribuição tomada como base, foi a variável do número de impressões, que também é a variável menos significativa semanticamente.

Para a taxa de cliques, que é a variável mais significativa, pois é um valor que independe da frequência de utilização do usuário, os resultados de alguns aplicativos (1, 3 e 5) foram bons, enquanto os demais foram acima de 1, o que é considerado ruim, pois significa que as distribuições a serem comparadas diferem da distribuição tomada como base.

Ao analisar a variável do número de cliques, é perceptível que esta foi a variável cuja distribuição para cada aplicativo foi mais distinta da distribuição base, pois os valores obtidos foram todos maiores do que 1.5, o que é considerado altíssimo, dado que valores próximos de 1 já indicam uma grande divergência.

Em geral, ao examinar individualmente os aplicativos, e levando em conta a relevância de cada variável, é possível dizer que os aplicativos 3 e 5 foram os que obtiveram um resultado mais parecido nas suas distribuições, quando comparados com as distribuições base.

O objetivo desta análise era observar as divergências, e, caso um aplicativo obtivesse uma grande divergência das distribuições base para as suas variáveis, classificar o aplicativo em questão como um suposto fraudador. Os resultados obtidos nesta etapa, numa análise superficial, sem levar em conta a semântica dos dados, poderia apontar os aplicativos 2 e 4, por exemplo, como possíveis fraudadores. Nestes dois aplicativos, principalmente no aplicativo 4, as distribuições das variáveis de taxa de cliques e número de cliques se afastaram bastante da distribuição base. Porém, ao analisar os valores destas variáveis comparados com os valores base, pode-se fazer uma análise mais profunda antes de classificar estes aplicativos como possíveis fraudadores.

Tabela 12: Comparativo dos valores do λ dos aplicativos com os resultados menos favoráveis e os valores base

	Taxa (λ) para a Taxa de Cliques	**Taxa (λ) para o Número de cliques**
Distribuição Base	0.778593182390314	0.155840917591323
Aplicativo 2	1.93187053267224	0.48661800486618
Aplicativo 4	3.35131576876541	0.765872711955273

A tabela acima mostra os valores do parâmetro da taxa da distribuição exponencial para as duas variáveis que obtiveram os resultados maiores para a divergência de Kullback-Leibler, a taxa de cliques e o número de cliques. Nesta tabela, foram dispostos estes valores tanto para os aplicativos com os resultados mais divergentes (2 e 4), tanto para as distribuições tomadas como base, com o propósito de comparar estes valores. É possível verificar que realmente os valores para os aplicativos diferem bastante do valor base, para as duas variáveis.

Porém, ao fazer uma análise levando em conta a semântica dos dados, é constatado que essa diferença pode ser considerada positiva. Dado que os dados analisados são a taxa de cliques e o número de cliques em propagandas, e que um suposto fraudador tem o objetivo de aumentar estas quantidades, para lucrar, uma distribuição fraudulenta seria caracterizada por mais usuários com valores altos para as duas variáveis. No caso destas variáveis, como a distribuição segue uma taxa exponencial, quanto maior o valor do λ, mais os dados estão

concentrados no início da distribuição, ou seja, a maior quantidade de usuários possui valores menores para estas variáveis.

Dessa maneira, levando em conta que o λ da taxa de cliques e do número de cliques para os dois aplicativos com a maior divergência são maiores do que o λ da distribuição tomada como base, é possível inferir que os dados destes aplicativos estão mais concentrados em valores menores, o que é positivo para o contexto estudado. Por isso, apesar de os aplicativos 2 e 4 terem obtido uma divergência alta dos valores bases, eles não podem ser taxados como candidatos a fraudadores.

O objetivo desta etapa foi tentar identificar se um dado aplicativo possuía dados muito diferentes dos considerados normais, porque isso poderia indicar que o *publisher* estava tentando fraudar cliques e impressões através de um grupo de usuários. Dados os resultados obtidos, e levando em conta a natureza dos dados estudados e das distribuições de cada variável, não é possível caracterizar nenhum dos cinco aplicativos analisados como um candidato a fraudador.

Caso um determinado aplicativo obtivesse uma grande divergência nas suas distribuições em comparação com as distribuições base, e se os valores dos seus parâmetros fossem menores do que os valores dos parâmetros das distribuições base, este aplicativo poderia ser um candidato a fraudador, pois, além da divergência, a análise semântica da divergência também faria sentido.

4.3. Detecção de *outliers* por valores espúrios

Outro experimento realizado no desenvolvimento do projeto foi o de detecção de *outliers* por valores espúrios. Esta técnica consiste basicamente em selecionar os valores extremos e caracterizá-los como *outliers*. No caso dos dados deste trabalho, como o objetivo é identificar fraudadores, os valores extremos são considerados os maiores valores, dado que a intenção de um suposto fraudador seria aumentar os seus números de cliques e de impressões. O objetivo deste experimento é caracterizar aplicativos e usuários como outliers estatisticamente.

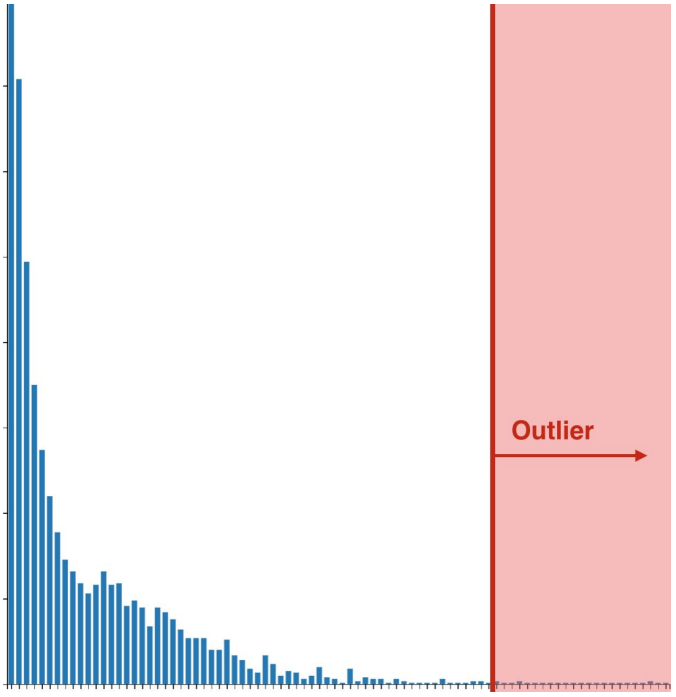
Este experimento consistiu basicamente em definir limiares para a caracterização dos *outliers* para cada variável e aplicar os valores definidos aos dados. Foram definidas duas categorias de limiares: estatístico e por especialistas. Além das categorias dos limiares, foram

analisados o conjunto de dados principal (dados agrupados por usuário) e o conjunto de dados por aplicativo (dados agrupados por aplicativos). Após a aplicação do método, foi feita uma análise levando em conta a relevância de cada variável analisada, e comparando os resultados das abordagens estatística e por especialistas.

4.3.1. Metodologia

De acordo com Aggarwal [23] uma forma popular de modelagem estatística em análise de *outliers* é a de detectar valores univariados extremos. Em casos como este, é desejável determinar valores de dados nas caudas da distribuição da variável em questão. A figura abaixo ilustra o princípio desta técnica.

Figura 12: Representação da técnica da detecção de *outliers* por valores extremos



Fonte: Autora

Aggarwal utilizou o termo "caudas" pois alguns tipos de distribuição possuem mais de uma cauda, como a distribuição normal, por exemplo. Como neste trabalho duas das variáveis seguem uma distribuição exponencial, e a outra segue uma distribuição lognormal, a cauda nestes casos serão os valores mais altos, como demonstrado na figura acima. Esta abordagem

irá selecionar *outliers* para cada uma das variáveis separadamente, posteriormente será feita uma análise global, compreendendo as três variáveis.

É importante ressaltar que a cauda em questão pode ter tamanhos variados, então é comum definir limiares para os valores antes de fazer a análise que irá rotular os dados como *outliers* ou não. No caso deste trabalho, foram propostas duas abordagens para os limiares: a abordagem estatística e a abordagem por especialista. Cada uma destas abordagens possui os próprios valores para os limiares utilizados, que serão apresentados nas subseções a seguir.

4.3.1.1 Limiares Estatísticos

A abordagem estatística é muito simples. A ideia é definir uma percentagem x e rotular como *outliers* os valores entre os $x\%$ maiores, para cada uma das variáveis analisadas. No caso deste trabalho, foi escolhido o valor de **5%** para todas as três variáveis analisadas: taxa de cliques, número de cliques e número de impressões.

Desta maneira, ao realizar a análise dos dados utilizando os limiares estatísticos, para cada observação, caso o valor de uma dada variável para esta observação estiver dentro do conjunto que compreende os 5% maiores valores para a variável em questão, esta observação será considerada um outlier para a variável analisada.

4.3.1.2 Limiares por Especialistas

A abordagem por especialistas é um pouco mais complexa, pois envolve a semântica dos dados analisados. A escolha destes parâmetros foi feita em conjunto com uma especialista, que lida diariamente com *publishers* de uma rede de anúncios. Por isso, ela conhece os valores que são considerados extremos para as variáveis.

A variável da taxa de cliques é uma ótima métrica para identificar *outliers*, pois se trata de uma percentagem, e independe da frequência de utilização do usuário. A média da taxa de clique varia muito de acordo com o formato do anúncio e com o tipo do aplicativo.

De forma geral, a média da taxa de cliques é de 1%, porém, a depender do tipo de anúncio e do tipo de aplicativo, valores até 5% podem ser considerados normais.

Para as demais variáveis, a imposição de um valor para o limiar não é recomendado. Pelo fato de o número de impressões e o número de cliques serem proporcionais à frequência de uso do aplicativo, além de dependerem da forma como cada aplicativo utiliza os anúncios, não é possível estabelecer um valor limite para uma observação ser considerada normal. Por isso, para estas variáveis, foi utilizada a mesma metodologia dos limiares estatísticos, mas considerando extremas as observações que fazem parte do conjunto dos **1%** maiores valores.

Desta maneira, para os limiares por especialistas, na variável de taxa de cliques, as observações que possuírem o valor da taxa de clique maior do que 5% serão consideradas *outliers*. Já para as demais variáveis, serão considerados outliers as observações cujo valor para a variável estiver dentro do conjunto que compreende os 1% maiores valores desta variável.

4.3.2 Resultados

Uma vez que foram definidos os valores dos limiares para cada abordagem, é possível gerar e analisar o resultado deste experimento. A análise dos resultados será feita em duas partes: uma analisando os dados do conjunto principal que é agrupado por usuário e a outra analisando os dados do conjunto agrupado por aplicativo. Cada uma das análises terá uma comparação das abordagens com limiares estatísticos e com limiares por especialistas.

Para a obtenção dos resultados foi elaborado um *script* que, de acordo com os limiares para cada abordagem, percorre os dados e gera um arquivo de extensão .csv. Este arquivo contém, para cada observação, valores representando a rotulagem da observação para cada variável. Esses valores são sempre 0 ou 1: 1 se for rotulado como *outlier* e 0 se for rotulado como normal.

Os resultados para uma dada observação contém três valores, cada um indicando se a observação é considerada um *outlier* com base em cada variável. Porém, cada uma das variáveis possui um grau de relevância diferente, por isso, com o objetivo de tornar a análise dos resultados mais coerente, decidiu-se utilizar pesos para as variáveis. A variável da taxa de cliques, por ser a mais significativa, tem um peso de 5 pontos no somatório da observação. Já a variável de número de cliques, que tem um valor semântico menor, tem peso 2, e o número

de impressões tem peso 1. Ou seja, cada observação terá uma pontuação em função das variáveis, que será utilizada para classificar a observação como um todo como *outlier*. Por exemplo, se uma observação tem os seguintes resultados: [1, 0, 1] (*outlier* para taxa de cliques, normal para número de cliques, e *outlier* para número de impressões), a sua pontuação será 6: $(1 \times 5) + (0 \times 2) + (1 \times 1)$. A tabela abaixo contém os significados de cada pontuação resultante dos pesos.

Tabela 13: Pontuações das observações e seus significados

Pontuação	**Tipo de Outlier**
8	*Outlier* para as três variáveis
7	*Outlier* para taxa de cliques e número de cliques
6	*Outlier* para taxa de cliques e número de impressões
5	*Outlier* para taxa de cliques
3	*Outlier* para número de cliques e número de impressões
2	*Outlier* para número de cliques
1	*Outlier* para número de impressões
0	Não possui *outliers*

Dado que cada variável possui um valor semântico, e que a taxa de cliques é a variável mais significativa, apenas as observações com pontuações mais altas foram consideradas *outliers* na análise final.

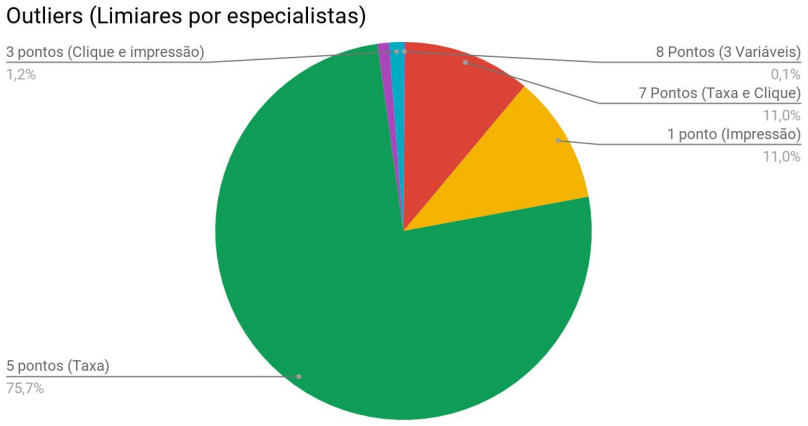
Por mais que a variável da taxa de cliques seja muito significativa, não é coerente rotular uma observação inteira como *outlier* porque ela se encaixa como outlier nessa variável. Uma dado usuário ou aplicativo pode ter tido 4 impressões de anúncio, e ter clicado em um deles. A taxa nesse caso seria 25%, mas não seria coerente classificar esta observação como um *outlier*, pois a observação tem dados poucos significativos. Além disso, quando um suposto fraudador tem interesse em fraudar, ele provavelmente o fará em larga escala, então, uma observação fraudulenta provavelmente seria *outlier* tanto na taxa de cliques quanto no número de cliques ou impressões. Por este motivo, na análise dos resultados serão consideradas *outliers* as observações com uma pontuação maior ou igual a 6.

4.3.2.1. Análise dos resultados por usuário

O conjunto de dados por usuário contém 10 mil observações. Na análise com base nos limiares por especialistas, 810 destas observações foram caracterizados *outliers* para pelo menos uma variável. Em contrapartida, na análise baseada nos limiares estatísticos esse número foi de 1135. Dados estes valores, é possível observar que a abordagem estatística rotula mais casos como *outliers*, e que a abordagem por especialista é mais seletiva na hora de considerar uma observação um *outlier*, o que já era esperado.

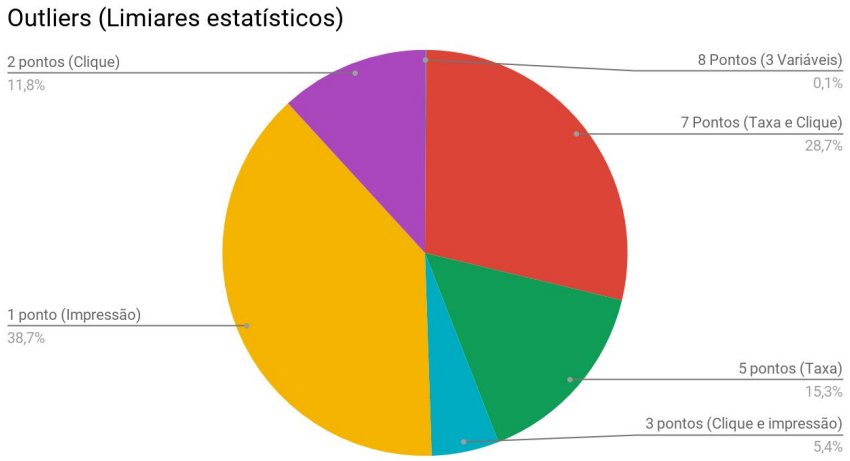
Os gráficos abaixo representam os resultados do experimento, e demonstram a proporção de outliers por categorias. O primeiro gráfico representa os resultados da abordagem por especialista, enquanto o segundo apresenta os resultados da abordagem estatística.

Figura 13: Gráfico dos *outliers* por usuário com base nos limiares por especialistas



Uma vez que os limiares por especialistas têm uma menor restrição para a classificação de *outliers* para taxa de clique (consideram qualquer observação com uma taxa acima de 5% um *outlier*), é esperado que o número deste tipo de *outlier* seja mais significativo. Em contrapartida, a restrição para as outras variáveis é maior, por serem considerados *outliers* os valores enquadrados nos 1% maiores para cada variável, por isso observa-se uma parcela menor com *outliers* para as outras variáveis.

Figura 14: Gráfico dos *outliers* por usuário com base nos limiares estatísticos



Uma vez que os limiares estatísticos possuem a mesma regra para a todas as variáveis, os resultados foram mais distribuídos entre as categorias. Pelo fato de serem classificados *outliers* valores contidos no conjunto dos 5% maiores, o número de *outliers* para as três variáveis foi maior do que a da abordagem por especialistas, e o número de observações com uma pontuação maior aumentou.

Dentre as 10.000 observações da amostra um total de 90 observações foram classificadas como *outliers* utilizando a abordagem por especialistas, e 326 foram classificadas como *outliers* para a abordagem estatística. Este resultado já era esperado, dado que a abordagem por especialistas é mais criteriosa de uma forma geral. Além disso, a abordagem por especialistas também possui um valor semântico maior. Por isso, apesar de os limiares estatísticos terem provido mais resultados, os limiares por especialistas são mais confiáveis.

Tomando como corretos os resultados da abordagem por especialistas, pode-se concluir que 90 usuários dentre os 10.000 analisados são candidatos a fraudadores. Este número totaliza 0,9% das observações, o valor foi considerado pequeno, dados os números de fraudes apresentados no capítulo 3. O próximo passo seria investigar mais de perto cada um destes 90 usuários, com o objetivo de comprovar que realmente está ocorrendo uma fraude.

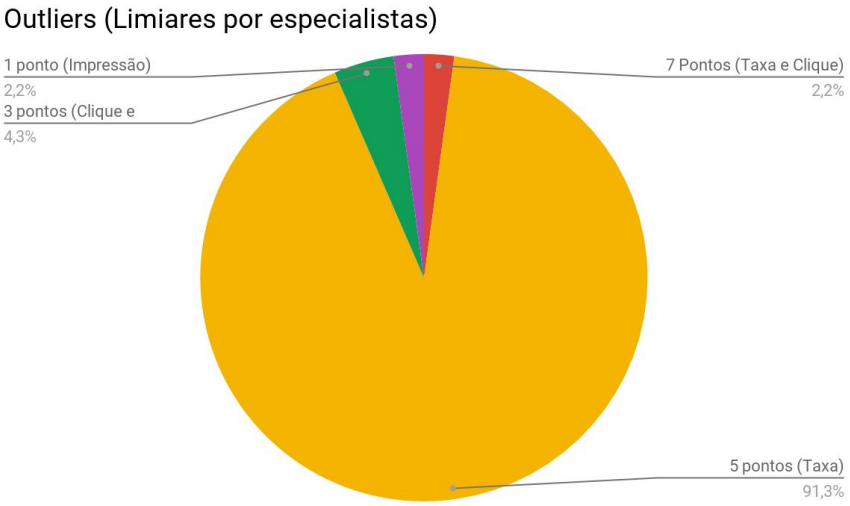
4.3.2.2. Análise dos resultados por aplicativo

O conjunto de dados por aplicativo contém aproximadamente 560 observações. Na análise com base nos limiares por especialistas, 92 destas observações foram caracterizados

outliers para pelo menos uma variável. Em contrapartida, na análise baseada nos limiares estatísticos esse número foi de 61. Dados estes valores, é possível observar que a abordagem por especialistas rotula mais casos como *outliers*, e que a abordagem estatística é mais seletiva na hora de considerar uma observação um *outlier*, o contrário da análise anterior, por usuário.

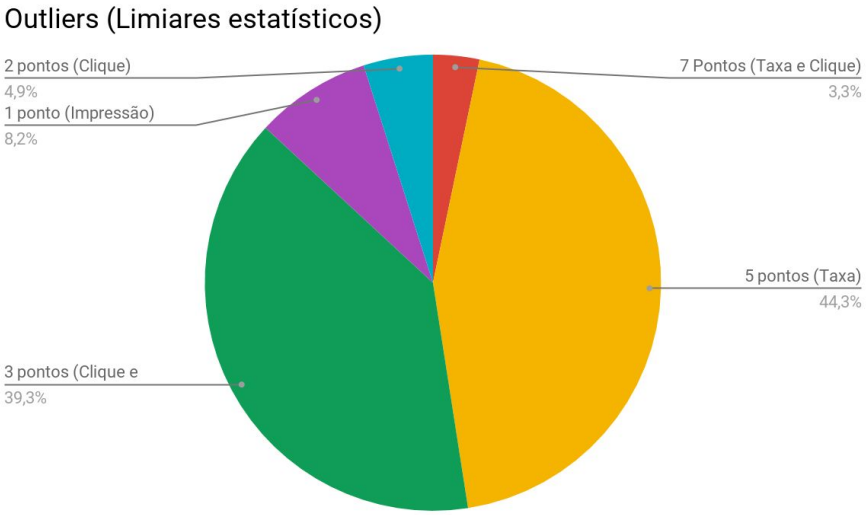
Os gráficos abaixo representam os resultados do experimento, e demonstram a proporção de outliers por categorias. O primeiro gráfico representa os resultados da abordagem por especialista, enquanto o segundo apresenta os resultados da abordagem estatística.

Figura 15: Gráfico dos *outliers* por aplicativo com base nos limiares por especialistas



A proporção dos resultados desta abordagem, assim como na análise por usuário, demonstram um volume grande de *outliers* para a taxa de cliques. No caso deste conjunto de dados isso era esperado, dado que o conjunto contém observações com poucos dados, resultando em muitos aplicativos com uma taxa de cliques maior do que 5%, que é o limite para esta variável.

Figura 16: Gráfico dos *outliers* por aplicativo com base nos limiares estatísticos



Já para a abordagem por limiares estatísticos, mesmo com o uso da mesma regra para todas as variáveis, os *outliers* por taxa de cliques e por número de cliques e número de impressões prevaleceu.

Considerando a pontuação, entre as 558 observações da amostra um total de 2 observações foram classificadas como *outliers* utilizando a abordagem por especialistas. Para a abordagem estatística, esse valor se repetiu. Porém, os 2 aplicativos classificados foram diferentes para cada abordagem. Por isso, é possível perceber que apesar de as técnicas terem tido um resultado final parecido numericamente, semanticamente eles variam.

Dado que a abordagem por especialista é mais confiável, é possível afirmar que entre os 558 aplicativos, 2 são candidatos a fraudadores. Este número representa 0,35% das amostras, e foi menor do que o da análise por usuário. Assim como na seção anterior, os próximos passos são investigar mais de perto os aplicativos em questão e tentar comprovar que eles realmente estão cometendo fraudes.

5. Conclusão

Dado que o problema das fraudes em anúncios *mobile* vem crescendo bastante, objetivo deste trabalho consistiu na tentativa de identificar possíveis fraudes em dados reais de uma rede de anúncios mobile. Para atingir este objetivo foram realizados experimentos envolvendo algumas técnicas estatísticas de detecção de outliers, já que um possível usuário ou aplicativo fraudador seria uma anomalia entre os dados.

Inicialmente, foi realizado um estudo sobre o funcionamento detalhado de uma rede de anúncios, com o objetivo de entender o motivo pelo qual as fraudes neste meio são tão comuns. Posteriormente, foi estudado sobre fraudes em geral e fraudes em anúncios mobile, além das técnicas de detecção de fraudes. Nesse estudo foi visto que é muito comum a utilização de detecção de *outliers* para a identificação de possíveis fraudes.

Após isso, foram realizados experimentos utilizando as técnicas estudadas. O primeiro experimento consistiu na análise da distribuição estatística dos dados, a análise foi feita de forma a comparar as distribuições de cada variável para cinco aplicativos com as distribuições consideradas base, que foram as distribuições obtidas com um conjunto de dados contendo observações de vários aplicativos. Apesar de ter havido discrepância entre as distribuições analisadas, não foi possível caracterizar nenhum dos aplicativos como um candidato à fraude, dadas a natureza e a semântica dos dados em questão.

O segundo experimento envolveu a detecção de *outliers* por valores espúrios, e foram utilizados dois tipos de limiares: os estatísticos e os por especialistas. Dentre os dois limiares, os por especialistas foram considerados mais confiáveis. Este segundo experimento foi mais bem sucedido, pois foi possível apontar alguns usuários e aplicativos candidatos a fraudadores.

5.1. Trabalhos Futuros

Por mais que este trabalho tenha explorado algumas técnicas de detecção de *outliers* em busca de encontrar fraudes em anúncios mobile, ainda existem muitos trabalhos que podem ser feitos. Na literatura não existem muitos trabalhos que abordam fraudes em

anúncios *mobile*, apesar de ser um tema bastante relevante. A maioria dos trabalhos sobre fraudes em anúncios aborda apenas os anúncios da *web*.

Trabalhos futuros nesta área poderiam envolver técnicas mais complexas de detecção de *outliers*, dado que as utilizadas neste trabalho foram as mais clássicas e simples. Uma outra abordagem interessante seria fazer uma análise multivariada, uma vez que este trabalho fez análises univariadas, apesar de ter explorado três variáveis. Também seria interessante a aplicação de técnicas envolvendo a distância entre os dados, dado que existem diversos algoritmos na literatura sobre técnicas envolvendo as distâncias. Além disso, poderiam ser aplicadas técnicas de aprendizagem de máquina para esta detecção, caso houvesse um conjunto de dados rotulado.

Outros trabalhos futuros interessantes seriam trabalhos para analisar mais de perto o comportamento de usuários ou aplicativos que são suspeitos de fraude. Outra iniciativa importante seria a de tentar identificar fraudes em tempo real, dado que não há muito a se fazer quanto o dinheiro perdido devido à fraudes, a não ser que essa detecção ocorra em tempo real.

Referências

[1] Kou, Y., Lu, C. T., Sirwongwattana, S., & Huang, Y. P. (2004). **Survey of fraud detection techniques.** In *Networking, sensing and control, 2004 IEEE international conference on* (Vol. 2, pp. 749-754). IEEE.

[2] **Bots Mobilize.** Disponível em: http://www.dmnews.com/ bots-mobilize/article/291566/. Acesso em: 30/10/2017

[3] Liu, B., Nath, S., Govindan, R., & Liu, J. (2014). **DECAF: Detecting and Characterizing Ad Fraud in Mobile Apps.** In NSDI (pp. 57-70).

[4] **Mercado Brasileiro de TI e Uso nas Empresas.** Fonte: Prof. Fernando S. Meirelles, 28ª Pesquisa Anual do Uso de TI, FGV-EAESP-GVcia, 2017. Disponivel em: <http://eaesp.fgvsp.br/sites/eaesp.fgvsp.br/files/noticias2017gvcia.docx>. Acesso em: 28/10/2017

[5] **Pesquisa sobre monetização da empresa Sweet Pricing.** Disponivel em: <https://sweetpricing.com/blog/2016/07/7-app-monetization-stats/>. Acesso em: 28/10/2017

[6] **Ad Example: Fun mobility Blog**. Disponível em: <http://blog.funmobility.com/2014/08/18/mobile-rich-media-ads-examples/> Acesso em: 20/11/2017

[7] Business Insider: **Explaining the mobile advertising system**. Disponível em: <http://www.businessinsider.com/explaining-the-mobile-advertising-system-2013-8> Acesso em: 18/11/2017

[8] **SDK Definition**: Gartner IT Glossary. Disponível em: <https://www.gartner.com/it-glossary/sdk-software-development-kit> Acesso em: 01/11/2017

[9] Acts, I. F. (2000). **University of Houston System Administrative Memorandum**.

[10] **Mobile ad fraud: What 24 billion clicks on 700 ad networks reveal.** Tune. Disponível em: <http://www.tune.com/blog/mobile-ad-fraud-24-billion-clicks-700-ad-networks-reveals> Acesso em: 30/10/2017

[11] De Cristofaro, E., Friedman, A., Jourjon, G., Kaafar, M. A., & Shafiq, M. Z. (2014). **Paying for likes?: Understanding facebook like fraud using honeypots.** In Proceedings of the 2014 Conference on Internet Measurement Conference (pp. 129-136). ACM.

[12] Forbes: **'Biggest Ad Fraud Ever': Hackers Make $5M A Day By Faking 300M Video Views**. Disponível em: <https://www.forbes.com/sites/thomasbrewster/2016/12/20/methbot-biggest-ad-fraud-busted/#fd9039a48990>. Acesso em: 06/12/2017.

[13] Motherboard: **Look at This Massive Click Fraud Farm that Was Just Busted In Thailand.** Disponível em: <https://motherboard.vice.com/en_us/article/43yqdd/look-at-this-massive-click-fraud-farm-that-was-just-busted-in-thailand > Acesso em: 22/11/2017.

[14] **UP TO 13% OF MOBILE VIDEO AD CLICKS ARE ACCIDENTAL.** Disponível em: <http://blog.pixalate.com/accidental-clicks-mobile-ads-data>. Acesso em: 10/12/2017.

[15] Hawkins, D. M. (1980). **Identification of outliers** (Vol. 11). London: Chapman and Hall.

[16] Hodge, V., & Austin, J. (2004). **A survey of outlier detection methodologies**. *Artificial intelligence review*, 22(2), 85-126.

[17] Stancescu, D. (2014). **Fitting Distributions to Dose Data.**

[18] Wessa, P. (2017). **Free Statistics Software, Office for Research Development and Education**, versão 1.2.1. Disponível em: <https://www.wessa.net/> Acesso em: 05/12/2017

[19] Kullback, S., & Leibler, R. A. (1951). **On information and sufficiency.** *The annals of mathematical statistics*, 22(1), 79-86.

[20] Joyce, J. M. (2011). **Kullback-leibler divergence**. In *International Encyclopedia of Statistical Science* (pp. 720-722). Springer Berlin Heidelberg.

[21] **Exponential Distribution**, Wikipedia. Disponível em: <https://en.wikipedia.org/wiki/Exponential_distribution>. Acesso em: 05/12/2017.

[22] Gil, M., Alajaji, F., & Linder, T. (2013). **Rényi divergence measures for commonly used univariate continuous distributions**. *Information Sciences*, *249*, 124-131.

[23] Aggarwal, C. C. (2017). **Probabilistic and Statistical Models for Outlier Detection**. In Outlier Analysis (pp. 35-64). Springer International Publishing.