Demystifying Local Business Search Poisoning for Illicit Drug Promotion

Peng Wang*, Zilong Lin*, Xiaojing Liao, XiaoFeng Wang
Indiana University Bloomington
{pw7, zillin, xliao, xw7}@indiana.edu

***Abstract*—A new type of underground illicit drug promotion, illicit drug business listings on local search services (e.g., local knowledge panel, map search, voice search), is increasingly being utilized by miscreants to advertise and sell controlled substances on the Internet. Miscreants exploit the problematic upstream local data brokers featuring loose control on data quality to post listings that promote illicit drug business. Such a promotion, in turn, pollutes the major downstream search providers' knowledge bases and further reaches a large audience through web, map, and voice searches. To the best of our knowledge, little has been done so far to understand this new illicit promotion in terms of its scope, impact, and techniques, not to mention any effort to identify such illicit drug business listings on a large scale. In this paper, we report the first measurement study of the illicit drug business listings on local search services. Our findings have brought to light the vulnerable and less regulated local business listing ecosystem and the pervasiveness of such illicit activities, as well as the impact on local search audience.**

I. Introduction

"OK Google, where to buy research chemicals?" Imagine that you seek from Google Assistant a nearby location to purchase research chemicals. Figure 1 shows the result returned by Google if you are in Denver, Colorado. The response is a structured "local business knowledge panel" recommending a "pharmacy" that is actually an online-only storefront (i.e., `apvpresearchchemicals.com`). Even worse, the store sells dozens of illegal addictive drugs such as fentanyl, methylone, and JWH-018! Such illicit local business listings are pervasive and are often recommended by leading local search providers such as Google, Bing, etc., as observed in our research (see Section VI), indicating that the underground drug promotions have already contaminated these providers' local business knowledge bases. Unlike blackhat search engine optimization, these promotions have circumvented the protections put in place by the providers, which often require local listings to go through restrictive vetting before getting into their knowledge bases, tend to be more prominent (among those highlighted few, with detailed information), and appear on multiple channels (search, voice and map). Never before has there been an effort to understand how the problem occurs or how serious the problem has become, and to develop effective mitigation of this emerging threat.

*The first two authors contributed equally to this project.

Network and Distributed Systems Security (NDSS) Symposium 2022
24-28 April 2022, San Diego, CA, USA
ISBN 1-891562-74-6
https://dx.doi.org/10.14722/ndss.2022.24284
www.ndss-symposium.org

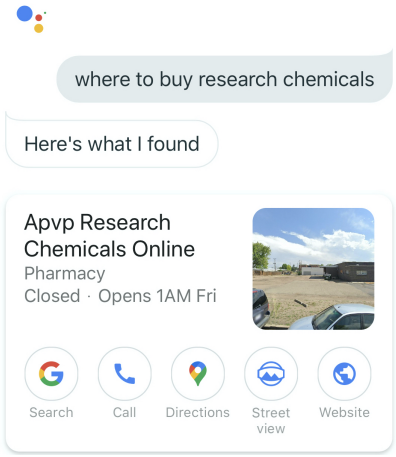

Fig. 1. Illicit drug local listing on Google Assistant.

Local business listing in jeopardy. Local business listings are business directory services widely offered by search engines (e.g., Google, Bing) and various local business portals (e.g., Yelp, Yellowpages). Through such a service, one can find an area store selling the searched-for product via the local knowledge panel returned by a search engine or the information labeled on a map (e.g., Google Maps, Apple Maps) or provided by various voice search systems (e.g., Google Assistant, Apple Siri). These local search services run on top of their knowledge bases that maintain structured information about local businesses, and are powered by a listing ecosystem in which various parties work together to create, collect, distribute, and display local business data to end users.

More specifically, listing providers collect data for their knowledge bases either from business owners' listing requests or through the purchase of business data from their partners. Verification of the authenticity of the information is provided by major providers such as Google and Apple: for example, Google sends business owners postcards to check their business addresses and makes phone calls to confirm their phone numbers. However, the quality of the data purchased from data partners tends to be difficult to control, relying solely on whether the partners have done their due diligence. This weakness is exploited by illicit drug promoters to contaminate knowledge bases, as discovered in our research.

Finding illicit drug listings on local search. In our research, we made the first attempt to understand the security impacts of the vulnerable listing ecosystem on local search services (i.e., local knowledge panel, map search, voice search). Our study

has led to the discovery of an emerging trend to promote illicit drugs through local business listings, with the promotion content entering the ecosystem from the problematic upstream—the local data brokers featuring loose control of data quality—and moving downstream to major search providers' knowledge bases. The information further reaches a large audience through localized-intent web, map, and voice searches. Note that even though illegal drug trading online has been studied before [55], [62], little is known about how the listing ecosystem is being abused to serve this underground business and the real-world impacts of these malicious activities, which we call ***I**llicit **D**rug **L**ocal **L**isting* or IDLL in our study.

Our study has been made possible through a new methodology for IDLL discovery and tracking, called IDLLSpread. To discover IDLLs, the approach utilizes the interconnections among listed drug-related businesses to discover those linked to IDLLs from a small set of known illicit online stores. More specifically, we found that illicit drug promoters often publish several listings with different names and addresses but share the same contact information, such as phone numbers and website URLs. In addition, different promoters tend to advertise the same set of illicit drugs that are often sought after by users. Such common promotion practices among the promoters connect different IDLLs, forming a graph, which enables us to find the unknown from the known listings through graph mining. Starting from a set of confirmed IDLL instances, IDLLSpread runs a semi-supervised label spreading algorithm on the graph to discover a larger set of illicit listings. Our study shows that this approach is highly effective, achieving a precision of 96.56% and a recall of 92.66%. Most importantly, using the IDLL instances we collected, IDLLSpread discovered 3,571 IDLLs from 94,856 collected local business listings. Given those IDLLs, we further designed an approach to determine their reachability on local search services. In particular, we proposed a method to automatically generate search queries reflecting innocent users' localized search intent and fed them into local search services. Among 8,546 search queries we generated, our approach reveals 4,176 (48.86%) of them associated with 1,689 IDLLs.

Measurement and discoveries. From the 3,571 IDLLs detected in our research, we found that such illicit listings are indeed pervasive: through `Yext.com` (a local listing scan tool [38] supporting search across 51 local data brokers), we extended these discovered IDLLs to 32,520 illicit listings, based upon their business names, phone numbers, and addresses. These listings promote various abusive drugs, including controlled substances like cocaine, fentanyl, heroin. They are quite successful in contaminating the knowledge bases of major local search providers (Google, Apple, Bing etc.). Specifically, they can be reached by 25.20%, 30.68%, and 2.73% of drug-related queries through voice, map, and web searches, respectively. From their drug listing graph, we identified 1,614 campaigns, with the largest one involving 962 IDLLs. Also discovered are the strategies for enhancing the visibility of IDLLs to end users, e.g., use of keyword stuffing ("Buy marijuana | cannabis oil | weed | CBD oil | kush buds") in business names and descriptions to attract search traffic, and inclusion of marketing words ("legal," "best," "online," "free shipping") together with locations to target those in certain areas. Another trick aimed specifically at voice search is to include question words, such as "where," "how," etc., and voice command words, such as "order," "buy," "get," etc., with business names. Of particular interest are the local addresses provided by the promoters: although they tend to be irrelevant or even nonexistent, some IDLLs hijack the contacts of legitimate local stores (e.g., Walgreens), using similar URLs and identical addresses to sell drugs without prescription.

Most importantly, we concluded that the security protection of today's local listing ecosystem is inadequate, allowing contaminated information to spread to prominent downstream listing services from the upstream data providers without proper vetting procedures in place. The problem fundamentally comes from the brokers that help promote local businesses to other listing services, which tend to be trusted by the services (such as search providers) but often do not enforce necessary listing policies. In our research, we investigated two of the most popular listing agents and several major local data brokers, and found that they all fail to properly verify the listing data submitted by untrusted parties, thereby opening the door of the whole ecosystem to IDLLs.

Contributions. Here we outline our contributions as below:

• *First study on IDLL*. We report the first systematic study on illicit drug local listings, a new kind of underground promotion that is becoming increasingly pervasive. Our study has been made possible through a new methodology that utilizes graph mining to infer suspicious local listings from known ones. Running our technique on 94,856 drug-related listings, we discovered 3,571 confirmed IDLLs, which have never been reported. We also discussed the strategies to mitigate such IDLL threats.

• *New findings and insights*. Our analysis on the IDLL instances has brought to light the impacts of this new kind of illicit promotion, not only on the audience of the conventional web search but also on those using map search and voice search, as discovered using a targeted voice crawler we built. Most importantly, through an investigation of the listing ecosystem, we present the evidence that today's listing ecosystem is less regulated and vulnerable. This new understanding contributes to better protection of the promotion model over various channels, including voice, which has never been studied before.

II. Background

A. Local Business Search

Searching for local businesses (a nearby restaurant, a local salon, a grocery store, etc.) has become increasingly popular, especially among mobile users. Major search providers, such as Google and Bing, support such queries through both text and voice channels, and return search engine results using local knowledge panels, map search results, or speech. Such results are generated using local business information (i.e., listings) collected from various sources such as web content providers (e.g., Wikipedia), local data partners (e.g., Yelp), or self-authoritative business owners. As mentioned earlier, our research focused on the abuse of local business search for illicit drug promotion, particularly the spread of IDLLs from data partners to local search services. The search services and data partners we studied are elaborated on below.

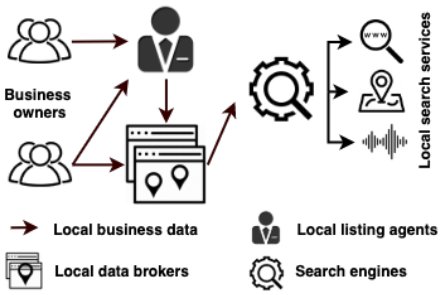

Fig. 2. Local business search ecosystem.

Local search services. We investigated three major types of local business searches: local knowledge panel, map search, and voice search, as illustrated in [23]. Note that although a general web search also returns localized search results, they are usually generated from web content instead of structured data collected from local data brokers, which therefore are beyond the scope of this paper.

• *Local knowledge panel*. Local knowledge panels are the information boxes that display the results returned from search engines (e.g., Google, Bing) for local queries, which contain business details such as location, contact, website, description, category, operation hours and reviews (see [23]). They are usually presented in the web-based search results when the businesses (e.g., restaurants) being queried appear in the search engine's local business knowledge graph [12], [41]. Such a knowledge graph utilizes the information from various sources. For example, Google reportedly purchases data from local data brokers Infogroup and Localeze [24]. A problem introduced by using such data is that their quality often cannot be assured. By comparison, the information directly submitted to the search engine by business owners often goes through a strict vetting process, involving address verification by postcards, email verification, phone verification, etc. [3], [33]. In our study, we analyzed the local knowledge panels on Google and Bing.

• *Map search*. Map services extensively serve local business searches as in [23], through which local business search results are displayed in a map view, including local entities' detailed information (e.g., address, website URL, phone number, review comments). Similar to local knowledge panels, map search also utilizes the local business listing data from local business data partners and its own knowledge graph [24]: for instance, Facebook, Foursquare, and Yelp are the data brokers for Bing Maps. To analyze the listings in map search results, we focused on the listings returned by the map services from the most popular web and mobile search engines, including Google, Apple, and Bing.

• *Voice search*. Local business search through voice assistants (i.e., VAs, like Apple Siri) is becoming more and more prevalent. It was reported that 58% of users find local business information via voice search [34]. Queries through VAs on mobile devices are localized to find nearby listings, with the listings usually returned in the form of structured Knowledge Panels (or Knowledge Cards, see Figure 1). Such knowledge cards are also powered by local business knowledge graphs. We studied the most common VAs on mobile devices, i.e., Apple Siri and Google Assistant in our research.

Local data brokers. Local data brokers (i.e., brokers) are the companies that specialize in collecting local business listings from a variety of data sources and selling such information to third parties [24]. Figure 2 illustrates the workflow of a typical broker. The broker first collects listings from various public (e.g., web content, government lists) and private sources (e.g., local business listing agents). It then performs data cleaning, de-duplication, and accuracy validation (e.g., phone call) before converting the listings into structured data. Such data is then sold to search providers to support their local search services through knowledge panels, map, and voice. Some of them (e.g., Yelp and TomTom) also provide other location-based services such as online rating and navigation. In our study, we looked into eight local data brokers including Foursquare, Yelp, Yellowpages, and MapQuest.

Note that local listing scan tools, such as `Yext.com`, will collect listing information from multiple local data brokers and provide search services. In our study, we used `Yext.com` [38], which aggregates listing information from its 51 partner brokers (e.g., Google Maps, Yahoo!), to further extend our findings. Also, it is the largest listing scan service provider in the industry [37]. To validate the reliability of `Yext.com`'s search results, we manually checked their existence on the partner brokers. The results show its veracity (see Section III-C).

Local business listing agents. Another important upstream source is *local business listing agents*, which act as an intermediary for local business owners to publish and maintain their listings on multiple brokers, helping the dissemination of their business information. We found the agents can also be abused to distribute illicit drug listings to brokers (see Section V-A).

B. Illicit Online Drug Promotion

Addictive and prescription drugs sold illicitly on the Internet is a serious public health threat. Under the FDA's regulations and the Ryan Haight Online Pharmacy Consumer Protection Act [39], it is illegal to advertise and sell controlled substances on the Internet in the U.S. For example, Google was fined $500 million for advertising online Canadian pharmacies to consumers [40].

To promote illicit drugs online, miscreants take advantage of many channels (e.g., email spam, search engine manipulation, social media) to attract users' traffic. For instance, it was reported that web search keywords were polluted and the rankings of web search results [62] were manipulated to promote illicit drugs. Besides, miscreants could publish and share the promotional information for illicit drugs through social media like Twitter and Instagram, which can reach a large number of users [59], [74]. Those promotions are usually run by the pharmaceutical affiliate programs of illicit drug merchants, which recruit publishers for drug advertisement.

To the best of our knowledge, we report the first systematic study on illicit drug promotion by poisoning local search results. Given that online users' search intents are increasingly localized, we believe that this emerging threat is in urgent need of serious attention and effective mitigation.

C. Adversary Model

We consider a miscreant to be a person who aims to pollute results returned from local search services (e.g., local knowledge panel, map search, and voice search) with illicit drug promotion, which violates those services' content safety policy [13]. For this purpose, the miscreant publishes illicit local business listings through the upstream brokers to spread the content to the local search services. We acknowledge that even under the stringent vetting procedures enforced by major local search providers (such as Google), direct contamination of their knowledge graph with illicit promotion content from the web (e.g., Wikipedia) or self-reported listings (e.g., Google My Business) can still happen. So, the IDLLs propagated from upstream brokers as observed in our study can only serve as a "lower-bound" for the pervasiveness of illicit drug promotion through local business search.

III. Methodologies and Evaluations

Our study aims at understanding the online illicit drug local listings on local search services (i.e., local knowledge panel, map search, and voice search). Figure 3 illustrates the pipeline of our methodology. By crawling a drug listing set with seed keywords, we built a graph mining-based method to discover IDLLs on the upstream local data brokers and then analyzed whether such IDLLs indeed affect search queries and results on the local search services.

A. Finding IDLLs on Local Data Brokers

Identifying IDLLs on data brokers is nontrivial: given the large number of listings on brokers, the portion of IDLLs is small and difficult to locate; so simply searching for drug names (e.g., Mephedrone) will return limited results or mostly legitimate listings because the miscreants usually use slang (e.g., research chemicals) in their listings. Also, compared with websites, IDLLs carry limited information (business names, website URLs, addresses, phone numbers, see Section II and Section III-C). Therefore, prior techniques [65] for analyzing the semantics of web content are less effective in this scenario.

Our methodology for detecting IDLLs is based on the observations that miscreants will post their IDLLs on many brokers with different business names but use the same contact information (e.g., phone number, website, address). This is because the miscreants will duplicate their listings on multiple brokers to make sure that the IDLLs reach downstream local search services. In the meantime, some miscreants will promote their IDLLs with different contact information. To conveniently manage their business and contact with customers, the promoters do not prepare the unique contact information (address, phone number, and URL) for each listing but instead reuse part of the information in their other IDLLs. This is similar to scam activities that only reuse phone numbers or email addresses [57], [67]. Leveraging the above observations, we identify IDLLs based on the principle of guilt by association (i.e., GBA) [72], using a graph mining algorithm to exploit the GBA hypothesis. Our algorithm, called `IDLLSpread`, takes the IDLL ground truth dataset and unlabeled listings (see below and Section III-C) as the input and outputs label score vectors to determine the label for each listing. When seeded with confirmed IDLLs, our approach reports the listings with high scores on the vector's IDLL label as IDLLs.

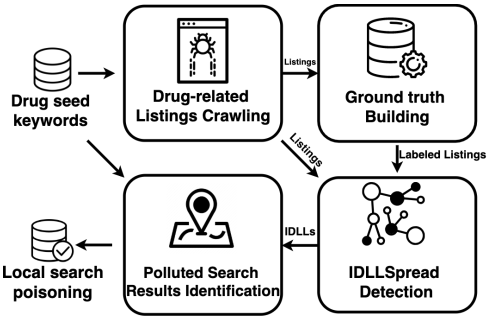

Fig. 3. Overview of methodology.

Undirected weighted graph construction. We represent the collected listings as a graph in which nodes describe listings and edges denote shared contact information (address, phone number, and URL), drug or slang terms, and "promotional terms" (or "selling arguments" such as "buy online") in listings' metadata. Note that we collected the "promotional term" list from related works on drug promotion [58]–[60] and further extended the list with their synonyms.

In our study, we define two types of edges: strong connection edges and weak connection edges. Different weights are assigned to different edge types. For the listings sharing the address, phone number, or URL (except for web builder sites), a strong connection edge with a large weight (100 in our study) is added between the nodes to represent the shared contact information, indicating that they belong to the same promoters and have the same label. If two listings contain the same drug name or slang term, a weak connection edge is added, indicating that they are possibly selling the same products and sharing the same class label. A weak connection edge is also used to connect the nodes sharing the same promotional term, which indicates similar promotional behavior. A relatively low weight value is given to the weak connection edge. The weights for the weak connection edges between two nodes sharing drug name, slang term, or promotional term are set to be 1, 0.2, 0.2, which achieves the best performance. If two nodes are connected by multiple edges, we only retain one with the largest weight.[1]

Such connection relations can be used to differentiate legitimate listings from illicit ones. More specifically, we utilize the listing metadata to link the collected listings to form a weighted graph `G = (V, E, W)`, where `V, E, W` are the nodes, the edges, and the edge weights, respectively. `V` is the set of listings, and `E` is the set of metadata (address, phone number, storefront URL, drug name, slang term, promotional term) shared among the listings. Each e in `E` can be represented by (u, v) where $u, v \in$ `V` are nodes, which indicates a certain relation between u and v. Each edge has a `w` determined by how the edges are connected. We are also given a training dataset, `L`, which consists of a set of labeled positive nodes, L_P (i.e., illicit listings) and negative nodes, L_N. The label of Node u is denoted by y_u, where $y_u = 1$ means that u is positive and $y_u = 0$ means that u is negative.

[1] We compared this approach with the model using the summed edge weight between nodes. It achieved a precision of 94.12% and a recall of 91.18% on the ground truth dataset with five-fold cross validation. The current setting had a better performance as shown in Section III-C.

GBA detection on weighted graph. To detect IDLLs, our idea is to leverage the connection among different listings to infer unknown labels from known ones. This label spreading is done using a semi-supervised learning algorithm, where nodes iteratively spread their current labels to neighbors, under the constraint that the ground truth nodes retain their initial labels. Based on our ground truth dataset, a label spreading algorithm is run on the graph to detect the illicit drug listings which are highly related to the existing illicit listings. The algorithm spreads the labels to other unlabeled nodes in the graph, in which the confidence of the IDLL label is determined by the node degree.

• *IDLLSpread algorithm.* IDLLSpread is designed based on the label spreading algorithm [75]. We first assign an initial labeling score $F_i = [f_{i1}, f_{i0}]$ to every node i in the graph, in which f_{i1} and f_{i0} are the positive score and the negative score, indicating the node's score to be positive or negative, respectively. For example, a node with a positive label will have $f_{i1} = 1$ and $f_{i0} = 0$ as the initial score. For an unlabeled node, the initial score is set to $F_i = [0, 0]$. The weight value w_{uv} of edge (u, v) describes the label closeness between u and v in classification. We set $w_{uv} = 0$ when u and v are not connected. Then, we leverage the label spreading algorithm to spread the initial labeling scores on the weighted graph to infer the labeling scores of unlabeled nodes.

In addition, although the edges constructed based upon popular illicit drug names or slang terms provide more evidence for the label spreading, they may introduce false positives to the listings in the presence of weak evidence. A low degree, few edge connections, and low-weight edges indicate the weak (or lack of) evidence for labeling a node. Thus, in our research, we leverage the confidence of the positive score based on the node degree to control false positives. More specifically, the confidence value will be used to punish the label score of the node with weak evidence to fine-tune the results. For the node with a lower degree, the confidence of its positive score becomes lower. The confidence θ is defined by Equation 1 in which d_i represents the degree of Node i, and d_{mean} represents the mean value of the degrees of all the nodes that only have weak connections. After each spreading iteration, the updated labeling score is corrected based on the node confidence calculated by Equation 2. The final label of Node i is determined by $\arg\max(F_i)$.

$$\theta = \begin{cases} \frac{d_i}{d_{mean}}, & \text{if } d_i < d_{mean}; \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

$$F_i = F_i \cdot \Theta, \qquad \Theta = \begin{bmatrix} \theta & 0 \\ 0 & 1 \end{bmatrix} \tag{2}$$

Filtering and extension. While IDLLSpread is used to discover the listings that sell controlled substances, the legitimacy of certain drugs varies under different jurisdictions. For example, marijuana can be legally sold by registered stores in a handful of U.S. states. To identify truly illicit listings, a set of rules (elaborated on later) are applied to filter suspicious IDLLs reported by IDLLSpread. In addition, we extend the results with a listing scanning tool, `Yext.com`, to check the appearance of IDLLs in other brokers.

• *Filtering*. To reduce false positives, we apply filtering rules on suspicious IDLLs selling marijuana and kratom, and their legality is determined by local legislation and approval. Specifically, even if marijuana and CBD oil are decriminalized or legalized in many states, all legal marijuana dispensaries are still under the approval and regulation of the state governments or related commissions, such as Oregon [29], [30], Florida [27], and Washington [26].

In our work, we used an allowlist collected from PotGuide [47], which aims to provide a complete directory of legalized marijuana dispensaries, to remove legal marijuana stores. Note that validating the correctness of PotGuide is challenging. In our study, we compared its allowlist with the three existing legal marijuana dispensary lists of Oregon, Florida, and Washington from government websites [26], [27], [29], [30]. The allowlists of these states provided by PotGuide turn out to be identical to those existing lists. The listings selling kratom in the states or cities where this drug is banned were also regarded as IDLLs [22].

• *IDLLs extension.* Our findings were based on the drug-related listing dataset collected from a set of popular brokers. To get a broader view of the IDLLs' impact on the entire listing ecosystem, we further extended our detection results with the listing scanning service from `Yext.com`. Given the name, address, and phone number of a listing, the service scans listing information from 51 broker partners (e.g., Bing Maps, Yahoo!, Facebook) to find out whether the listing has been published on them. With the help of this service, we acquired more details about the detected IDLLs, such as other brokers they affect, their alternative business names and phone numbers.

Although the extended listings share the same phone numbers or addresses as the input listings, they might not relate to the illicit ones we found, since legitimate parties' contacts could be stolen by the IDLLs. To reduce false positives in the extended listings, we constructed an undirected, unweighted graph in which the extended listings and the detected IDLLs are connected to determine their overall similarity. Different from the undirected weighted graph for IDLLSpread, this graph is unweighted and the nodes (i.e., listings) are connected by addresses, phone numbers, and URLs. Specifically, we used Jaro Similarity [20] to measure the closeness between each extended listing and the IDLLs in the cluster (see Section IV-B), and labeled it as an IDLL when the similarity (with at least one IDLL) was above 0.85.

B. Identifying Polluted Local Search

To understand IDLLs' impacts on search engines and users, we studied how they affect the search results reported by local knowledge panels, map search, and voice search. For this purpose, we developed and ran a methodology that automates query generation for local search services, and further gathered and analyzed the local search results produced by the queries.

Query generation. From the drug seed keywords (see Section III-C), we generated a large set of queries frequently searched by online users through different search interfaces, as elaborated below. Particularly, since users are more likely to use natural questions when interacting with VAs, we also collected drug-related questions as the voice search queries.

• *Autocompletes.* Given a drug keyword (e.g., "cocaine"), autocompletes [11] from the search engines provide the alternative popular queries (e.g., "cocaines drug effects") made by the users. Also, to get the question autosuggestions (e.g., "how cocaine is made"), we added the question words (e.g., "how," "what," "does") as the prefix of drug keywords. In our study, we fetched the autocompletes from Google for popular queries from users.

• *Keyword tool.* TextOptimizer [50] is a popular keyword tool that provides the most frequently asked questions (e.g., "what is cocaine") about given topics (e.g., "cocaine"), reflecting the users' search intentions. We fed the drug keywords into the tool and fetched the questions.

• *Related questions.* Search engines also list variants of a given natural language queries within "People also ask" (i.e., PAA or Related Questions) [9] in search results. We crawled the questions in the PAA section of drug keywords on Google.

• *Location.* Local business search results are closely related to geolocations. To cover the results as broadly as possible, we appended the queries with 50 different locations in the U.S. (the largest city of each state). Search results for such queries would be localized to the corresponding locations.

Query through local search services. Further, we measured how the detected and extended IDLLs from brokers can be reached by local search users.

• *Local knowledge panel and map search.* Given an IDLL, we first extracted the drug keywords used for the drug promotion. A list of queries were then obtained from the aforementioned query generation approaches. For local knowledge panels, we used Google and Bing web search, and parsed the knowledge panel results. For map search, map services from Google, Bing, and Apple were queried. For knowledge panel information, we directly searched on each service interface and crawled the authentic search results. For map search results, we used the map API on each local search service.

• *Voice search.* We set up a voice search crawler pipeline to gather voice search results. More specifically, to interact with the real VA systems, the generated queries were first synthesized to voice commands. Then we fed the commands to VA systems, collected the search results, and checked how the IDLLs impact the search of VAs. In our study, we used Amazon Polly Text-To-Speech [2] to synthesize the voice commands (in English). The voice commands were fed to the most popular VA products on smartphones [44]: Apple Siri and Google Assistant. Before querying each voice command, we simulated the voice search process and played a trigger command (e.g., "Hi Siri") to wake the VAs into the listening mode. We specified a time interval (i.e., 22 seconds in Apple Siri and 15 seconds in Google Assistant) between commands for the VAs to recognize the commands and respond with search results. Videos were recorded for the interaction process, and we took a screenshot image once per time interval as the search result. Further, we leveraged the Google OCR API [14] to recognize the content on the images. We elaborate the measurement study of user impact on VAs in Section VI-A.

C. Implementation and Evaluation

Implementation. In our implementation, we utilized the `NetworkX` library [54] to construct the undirected weighted graph. Then, we built the IDLLSpread algorithm by combining the classification module of `NetworkX` with the confidence computing module. We ran our implementation of IDLLSpread on an R730xd server with 39 Intel Xeon E5-2650 v3 2.3GHz, 25M Cache CPUs and 256GB memories. Web and map searches on Apple, Google, and Bing were recorded on a laptop (HP Pavilion 15t). Voice searches on Apple Siri and Google Assistant were tested on an iPhone XR.

Datasets. Similar to the evaluation reported by prior work [61], we evaluated IDLLSpread over the collected 94,856 listing set, including labeled ground truth dataset and unlabeled listings.

• *Drug seed keywords*. In our study, we utilized the names of commonly abused illegal and controlled substances as drug seed keywords. These names are reported by the National Institute on Drug Abuse (NIDA) and the Drug Enforcement Administration (DEA), which are a ready reference for law enforcement personnel to identify and control drugs [43], [45]. The seeds contain the drug's international nonproprietary name (e.g., oxycodone, cocaine) and their brand names (e.g., OxyContin, Percocet), which we called as drug names. Also, the drugs' common "street" or "slang" names (e.g., oxy, white powder) are included in the aforementioned data sources. In total, we collected 1,850 drug seed keywords, including 759 drug names and 1,091 slang terms.

• *Drug listing set.* To understand how the local search results are polluted to promote illicit drugs, we collected the drug-related listings from the local listing ecosystem. For this purpose, we searched the 1,850 drug seed keywords on the popular brokers [42], [46], [51], [52]. Specifically, we targeted the eight major brokers (i.e., Factual, Foursquare, Infogroup, Localeze, Manta, MapQuest, Yellowpage, and Yelp)

Using the drug seed keywords, we crawled the related listings through brokers' search interfaces, which require keywords and locations as inputs. To optimize collection coverage, each seed was searched in the largest city of each U.S. state. In total, we discovered 94,856 drug-related listings that have drug keywords in their metadata (name, description, category, etc.). Each listing carries the business name, website URL, address, phone number, description, and category. Figure 4 shows the amount of collected data on each local data broker.

• *Ground truth.* To build the ground truth datasets, two security professionals spent eight days on manual validation. A case was flagged when both annotators reached an agreement. Here, we set the inter-coder reliability measured with Cohen's kappa coefficient [68] to $\kappa = 0.91$. We have released this annotated dataset [23].

- *Badset.* We used a similar approach as the ones proposed by previous studies [58]–[60], which look for promotional terms and drug keywords in tweets to find illegal drug stores, to identify IDLL candidates with clear drug promotion signals from the drug listing set. Specifically, we first identified 1,998 IDLL candidates whose metadata contain promotion signals and further manually checked their illegitimacy. To ensure

the quality of this dataset, we utilized promotion signals in the form of promotional term and drug name combinations. Such combinations often appear in illicit drug promotion as mentioned in the prior research [58]–[60], such as "order D online," "the best D," "legal D near me," etc. D is one of the 759 seed drug names. The list of promotional terms includes 58 keywords (e.g., "buy," "online," "bitcoin," "the best," "show me," "close to me"). We also released these keyword lists [23].

We further manually checked IDLL candidates' existing websites and context to find out whether they were indeed illicitly promoting drugs. In particular, we inspected each site's webpage including its content and drug selling list. We also checked the presence of payment and purchase processes, etc., to find whether the site was selling illicit drugs or linked to other sites. If no such website was observed, we inspected the listing context to understand its semantics to determine whether it was illicit (e.g., "buy research chemicals and bath salts online") or not (e.g., "marijuana card doctors"). In the end, we identified 1,718 suspicious IDLLs as the badset.

- *Goodset.* To get the good set, for each drug seed keyword, we randomly selected three listings containing this keyword from the collected drug listing set, and 5,105 of them (among 5,550 selected listings) were further manually verified to be legitimate. We also included legitimate local entities in the goodset from the National Directory of Drug and Alcohol Abuse treatment facilities [49] and SAMHSA Opioid treatment program directory [48]. For IDLLSpread, only the ones related to the drug listing set were chosen, including 292 listings. In this way, we collected 5,397 accredited good listings.

Result and evaluation of IDLLSpread. We evaluated the model over the ground truth dataset, using five-fold cross validation. Our approach achieved a precision of 96.56% and a recall of 92.66% on average, after its convergence (after 72 iterations on average). Without the knowledge about the ratio of bad listings in the real world, we further evaluated the model over the validation sets with various bad listings ratios. As shown in Table I, although both precision and recall changes under different bad listings ratios in the validation set, the accuracy is stable across all the settings. Looking into false positive listings, most of them are not for drug trading but include drug names or slang terms in their profile (e.g., "cannabis investing spot" for cannabis farm investment). Also, most of the false negatives carry only low frequent drug names or slang terms (e.g., "green peace") that renders IDLL labels hard to propagate from positive nodes to these listing nodes.

To check the model's generality, we selected the ground truth with the geolocation of Florida as the training set containing 174 good listings and 120 bad listings, and those from Texas as the test set (including 150 good listings and 97 bad listings). Running our model on this dataset, it achieved a precision of 90.91% and a recall of 92.78%, which indicates the good generality of the model. Our tool identified 7,509 to be suspicious IDLLs, including 1,718 in the ground truth set. The detection results on the drug listing set were further inspected manually on 500 sampled listings, and our approach achieved a precision of 97.14% on the drug listing set.

We also found that weak connections are effective in detecting IDLLs, making an important contribution to finding new IDLLs. Using the undirected weighted graph model, we observed that only 38.5% of the listings in our graph have strong connections by sharing contact information. Additionally, in the ground truth badset, 769 listings also do not have any strong connections with other labeled or unlabeled listings. Further, among the 7,509 suspicious IDLLs, 4,920 do not have strong connections with labeled or unlabeled listings and are therefore identified by weak connections. We also experimented with a model which we removed all the weak connections. According to the five-fold cross validation, this model achieved a 99.96% precision while only a 11.46% recall. On the other hand, we got a more balanced result of precision (96.56%) and recall (92.66%) with the model utilizing both strong and weak connections, indicating that weak connections can help reduce the bias in detection results.

TABLE I. IDLLSPREAD PERFORMANCE WITH VARIOUS RATIOS OF BAD LISTINGS

Ratio	5%	10%	20%	30%
Accuracy	98.71%	98.37%	97.73%	97.06%
Precision	83.32%	90.86%	95.27%	97.92%
Recall	93.44%	93.12%	93.28%	92.18%

TABLE II. IDLLSPREAD PERFORMANCE ON THE VALIDATION SETS CONTAINING THE FIRST TYPE OF COUNTERMOVES

Ratio	One metadata replacement			Two metadata replacements		
	30%	60%	90%	30%	60%	90%
Accuracy	93.33%	90.51%	82.09%	92.13%	85.37%	82.01%
Precision	88.14%	82.96%	75.60%	88.13%	74.67%	74.56%
Recall	83.51%	76.19%	40.14%	75.62%	60.52%	38.35%

In our study, we also evaluated the model effectiveness against simple countermoves. Specifically, we considered two kinds of countermoves: (1) miscreants camouflage bad listings as good ones by using the legitimate listings' metadata; (2) miscreants update their slang terms with new words (See Section IV-C). To craft the first type of countermoves, we replaced one or two elements in each bad listing's metadata (address, phone number, or URL which are used for strong connections in the graph construction) with the ones from a random legitimate listing in the ground truth dataset. Besides, we tested the model under different ratios of the listings involving this countermove in the bad listing data of the validation sets. When constructing the second type of countermoves, we updated thirty slang terms using by IDLLs with a set of seventeen newly-discovered drug slang terms [76]. Note that as newly emerging slang terms, these words do not exist in the bad listings of the training sets.

Given these two countermoves, we evaluated the model's effectiveness in term of precision, recall, and accuracy, using five-fold cross validation. As shown in Table II, the increase of the first countermove ratio led to a drop in the effectiveness. Meanwhile, our model revealed its capability against countermoves. Even when 60% of the bad listings were countermoves, not many of them evaded the detection (with the precision and the recall both still above 60%). The accuracy was always above 80% across different ratio settings. In the evaluation to the second countermove, the slang term refreshing did not affect much on the model performance, which achieved a precision of 95.71% and a recall of 92.44%. This is because the countermove invalidates only one weak connection after updating one slang term, and other strong or weak connections are still effective in label propagation.

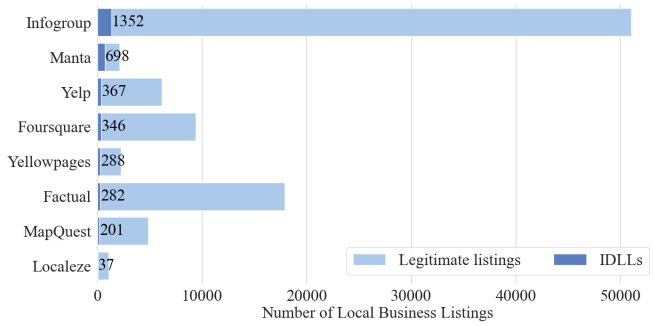

Fig. 4. Numbers of IDLLs on different local data brokers.

Result and evaluation of filtering and extension. IDLLSpread reported 7,509 listings from the collected brokers, among which 3,571 IDLLs were confirmed after filtering. Using these detected IDLLs, we further identified 45,169 extended listings at the extension step. After computing similarity, we obtained 28,949 extended IDLLs across 51 brokers in the ecosystem. We further randomly sampled 500 extended IDLLs for manual inspection to check whether they were indeed IDLLs, concluding that the `Yext.com` extension achieved a precision of 97.60%.

We also found that the search results of `Yext.com` are indeed reliable: that is, all the results it returns are public listings provided by its brokers, which are indeed visible to the users of the brokers (such as Yelp), so any IDLL included will also be displayed to the users. More specifically, we randomly sampled 10 IDLLs extended by `Yext.com` through each broker, and looked them up directly in the broker's own search engine (e.g., Yelp); we observed that 95.88% of the IDLLs could still be found in the search results.

IV. Understanding Illicit Drug Local Listings

In this section, we report a measurement study on 32,520 detected and extended illicit drug-related listings discovered by IDLLSpread to understand how local search results were polluted by illicit drug promotion.

A. Overview

Scope and magnitude. Our study reveals the pervasiveness of this emerging threat, with 3.76% of the 94,856 collected listings found to illicitly promote drugs. Figure 4 illustrates the number of the IDLLs on different brokers. Among them, Infogroup was mostly targeted (37.86%), followed by Manta (19.55%) and Yelp (10.28%). Among all 32,520 detected and extended IDLLs, "marijuana," "CBD oil," "research chemicals," and "steroids" were the four most popular abusive drugs keywords. These drugs had been promoted by 78.08%, 6.75%, 5.52%, and 2.97% of IDLLs, respectively. Also, 18.90% of the IDLLs advertised more than one type of drugs. Table III lists the top 10 drug categories promoted by the IDLLs which cover a wide range of controlled substances and prescription drugs.

Slang for promotion. It is interesting to note that slang terms (e.g., "research chemicals" as in Table III) were frequently used for drug promotion. In 32,520 detected and extended IDLLs, 7,466 (22.96%) IDLLs contained slang, in which 3,874 used slang terms in their business names, and 4,445 included slang terms in other metadata such as descriptions and categories. The most popular slang terms included "research chemical(s)," "edibles," "420," "mmj," etc., as shown in Table IV. Most of the popular slang terms refer to marijuana.

TABLE III. Top 10 abusive drugs promoted by IDLLs

Drug	# of IDLLs	Drug	# of IDLLs
marijuana	25,393	kratom	840
CBD oil	2,196	painkiller	596
research chemical(s)	1,796	bath salt(s)	569
steroids	967	Cialis	463
Viagra	942	oxycodone	407

IDLL storefront characterization. To understand the miscreants' behaviors, we further analyzed their website content. In our dataset, we found 1,210 distinct storefront URLs, in which 960 (79.3%) were still alive, and we crawled the content on the websites. For those which were not accessible at the time of our investigation, we used WayBack Machine Scraper [35] to obtain their web content.

On such web content, we analyzed how the owners of IDLLs arrange payment and drug delivery. Interestingly, multiple storefronts use worldwide shipping and Bitcoin payment as their featured services to attract potential buyers, as promoted in their business profile. More specifically, we found that 44% of the storefronts deliver nationwide in the U.S., and 15% of them provide worldwide shipping. Although mailing such controlled substances is illegal, the storefronts claimed to use mainstream shipping services such as USPS, FedEx, and UPS. We also checked how the IDLLs process the transactions. It turned out that 27% of them accept Bitcoin, and 39% accept payment through banking systems via credit cards, checks, and direct bank transfer. It was observed that 20% of the IDLLs leverage online payment instruments such as Paypal, CashApp, and GreenBeanPay. Money orders and gift cards are used by 27% of the storefronts. During our investigation, we found that for traditional payment instruments (e.g., wire transfer), sellers share account numbers to allow transactions to proceed only after their customers proactively initiate the conversation with them and place orders.

We also observed that IDLLs' illicit drugstore websites enjoy a long lifetime. As mentioned, 79.3% of the collected storefront URLs were still available to visit at the time of our study. To estimate their lifetime, we checked the WHOIS information of these domains. For those unavailable domains, we checked their snapshots on Wayback Machine and regarded the period between when Wayback Machine first and last indexed the websites as their lifetime. In total, we obtained the lifetime information of 1,092 unique websites with a median of 0.57 years lifetime and 8.52% of them live for more than three years. In addition, the number of new illicit storefronts has been increasing since 2008.

IDLL phone number analysis. In all the obtained IDLLs, 2,852 unique phone numbers make important connections between storefronts and users. Using `everycaller.com` and `spamcalls.net`, we found that 26 phone numbers belonging to 106 IDLLs were marked as spam numbers. We also checked `twilio.com` for each phone number's profile and found the numbers of the VoIP, mobile, and landline numbers are 957, 474, and 1,186, respectively. The type of the rest of

TABLE IV. TOP 10 DRUG SLANG TERMS IN IDLLS

Drug	# of IDLLs	Drug	# of IDLLs
research chemical(s)	1,796	top shelf	289
edibles	1,513	syrup	280
420	1,145	sweet leaf	127
mmj	730	mile high	104
bath salt(s)	569	mary jane	102

235 phone numbers is unknown in `twilio.com`, which are probably fake or foreign numbers. In general, IDLLs tend to choose the VoIP or landline numbers as their contact numbers. We also notice that promoters for different drugs have different preferences over the type of phone numbers: for the drugs under loose regulation such as marijuana, Viagra, and Cialis, the promoters tend to use landline numbers indicating they may have physical stores. By contrast, for strictly controlled drugs like research chemicals and oxycodone, their promoters tend to choose the VoIP phone to evade detection.

B. Discovering Illicit Drug Campaigns

It's important to understand the relations among the IDLLs to demystify IDLL campaigns. In this section, we illustrate our studies on these campaigns, which were polluting the local search results for illicit drug promotion. We first introduce our methodology and then describe our findings in detail.

Methodology. After applying IDLLSpread on the collected listings, the obtained IDLLs which are related with others will be connected as IDLL clusters on the graph. The IDLLs in such clusters sharing strong connections (i.e., phone number, website URL) are believed to belong to the same campaign. As such, the IDLLs were grouped into 1,614 clusters and 1,463 unconnected nodes; 15 clusters have a size of more than 100. The largest cluster contains 962 IDLLs. This cluster focused on selling research chemicals, bath salts, and marijuana. For the IDLLs in this cluster, there were 65 unique IDLLs with different combinations of names, addresses, phone numbers, and URLs. The rest were the same IDLLs in different brokers.

Discoveries. We further analyzed the details of the IDLL clusters from their metadata such as URLs, phone numbers, and business names. Below, we demonstrate our findings.

• *Consistency of URL and phone number.* Website URLs and phone numbers are the key information in the IDLLs, which connect the miscreants and potential buyers. We observed that most of the listings in one campaign have the same URLs and phone numbers. Specifically, 1,300 (80.55%) of the campaigns used only one phone number among all their listings, and 1,537 (95.23%) of them used the same online storefront.

• *Location spam.* IDLL campaigns usually only operate online with no physical storefronts. To be reached by potential buyers from different locations, miscreants always listed their IDLLs at multiple locations (different cities or countries). Only 178 (11.03%) campaigns registered their listings at one address.

• *Naming patterns.* When looking into the IDLLs in one campaign, different business names were frequently used for drug promotions. We summarized the naming patterns used by the campaign owners as follows: **(1)** to attract users with different drug search intents, the owners would create several listings

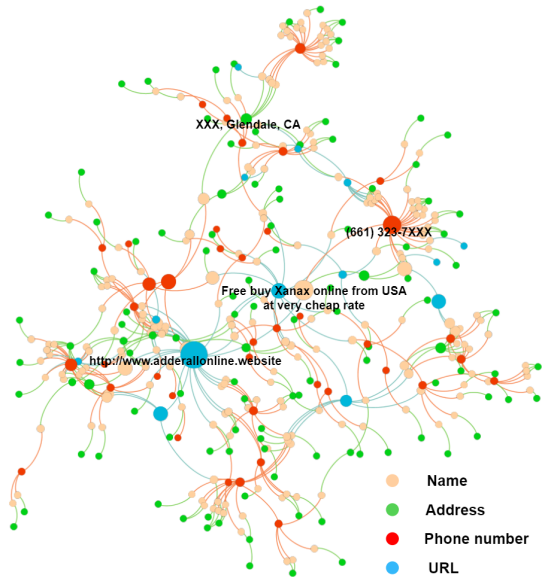

Fig. 5. The subgraph of a campaign with 207 IDLLs.

with different drug keywords in the business names, (e.g., "Order *Xanax* Online in Mississippi and Get Free Delivery" and "Order *Roxycodone* Online in Mississippi and Get Free Delivery"); **(2)** to promote the listings more widely, the owner would combine the same name with different locations (e.g., "Kratom *Chicago*," "Kratom *Columbus*," "Kratom *Seattle*"); **(3)** the campaigns often use different promotional terms in the names (e.g., "*Buy* steroids *with credit card*," "*Top* steroids *online*"); **(4)** completely different names may be used in the same campaign (e.g., "Mydrugpill," "Buy Tramadol Online").

Case study. One of the largest campaigns we discovered was illicitly promoting prescription drugs (e.g., painkillers, Roxycodone, Xanax), which consisted of 207 IDLLs (75 unique ones) on various brokers. The connections among the campaign are shown in Figure 5. Besides the nodes representing the IDLLs, the shared contact information of these IDLLs (addresses, URLs, and phone numbers) also display as nodes in the graph to reveal the relationships in this campaign. Interestingly, the figure clearly shows that the IDLLs in the campaign are not linked by only one type of nodes: the top 10 centroid nodes include four business names, four URLs, and two phone numbers. With the help of the graph structure, we can reveal the hidden relationships among IDLLs which seem to be irrelevant but belong to the same illicit campaign.

Looking into the IDLLs names, 138 IDLLs had various promotional terms in their names, like "how to use," "without prescription," or "overnight delivery." Among these listings, 118 of them used the pattern "buy `D` online" (`D` is a drug name) in their business names. To reach the buyers at different locations, 71 of IDLLs added locations in their names, such as "buy `D` online `L`" (`L` is a location, such as "Mississippi," "Alabama," or "Puerto Rico" in the U.S. and even foreign locations like the U.K. and Canada). The claimed IDLLs' addresses in this campaign leaned to be located at big cities (e.g., "New York," "Chicago," "Philadelphia"). Taking advantage of the huge population of big cities, such addresses would attract more buyers to visit the IDLLs there.

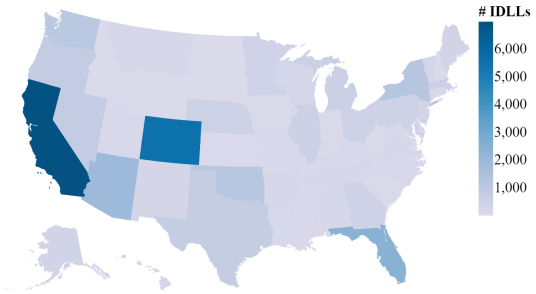

Fig. 6. Geolocation distribution of IDLLs in the U.S.

C. Promotion Strategies in IDLLs

We found that IDLL miscreants targeted legitimate drug listings with phishing attacks. The miscreants also tend to utilize blackhat SEO techniques, such as location stuffing (i.e., adding unrelated location information) and query spam (i.e., using questions as the listing names), to promote the listings.

• *Local listing phishing.* Most interesting is the discovery that some IDLLs even used the same addresses as the legitimate listings. When comparing the IDLLs with those listings in the allowlist (see Section III-A), we found that 34 IDLLs used the same addresses as the legitimate ones but featured different websites and phone numbers. For example, an IDLL "Buy Drugs Online" was located at the same address as the legal pharmacy store Walgreens, but it used a phishing URL (`www.walgreens-pharmacy.com`) to deceive online users. Another example is the IDLL "Weed For Sale, We do Delivery All Country" (`ww1.coloradogreendragon420.com`) which had the same address as a legal marijuana dispensary but differed in other metadata. Such listings will mislead the local search users and take search traffic away from legitimate listings.

• *Location abuse and stuffing.* Location is an important factor for drug promotions and has been abused by miscreants. To be searched by users in a wider area, some listings' profiles contained different cities or country names. Specifically, many illicit listings explicitly used big city names in their business names, and had their addresses located at big cities, such as "Best Weed Delivery *Los Angeles*." Country names also appeared in the listing names to promote their business with worldwide search intents. For example, an IDLL is named as "Buy Marijuana Online *USA*, Buy Weed Online *UK*, Buy Marijuana Online *Australia*." On the other hand, for voice search users, the location information is implied in the queries when searching through VAs. This will make the listing with a location in the name achieve higher visibility than others. Figure 6 demonstrates the geographical distribution of the IDLLs in the U.S. We can observe that 6,990 of IDLLs were located in California, the state which has the most IDLLs, followed by Colorado (5,534) and Florida (2,449). When looking into the cities, Denver was the most popular city the illicit listings targeted, followed by Los Angeles and Phoenix.

• *Questions as business names.* Many of the illicit local business listings used questions in their business names. This approach took advantage of the fact that voice search users always use natural questions to make queries on VAs. So, using question keywords directly as the listing names makes the listings easily match the voice queries and unfold in the search results; 1,826 IDLLs used questions in their listings, in which 42, 1,775, and 14 IDLLs used questions in their store names, descriptions, and even URLs, respectively. We noticed that the stores using questions as names are more accessible to voice search users (see Section VI).

TABLE V. LOCAL BUSINESS LISTING AGENTS

Listing agent	Operation time	Cost (/year)	# of partner brokers	Submission
BrightLocal	∼1 month	∼$100	20+	Manual
MozLocal	∼3 weeks	∼$200	20+	Automated

• *Keyword stuffing.* Keyword stuffing is a common blackhat SEO method that stuffs popular keywords into the promotion targets (e.g., webpages), aiming to lure users' search traffic. We observed that 1,651 IDLLs used this approach with two common practices. First, the listings directly displayed a list of drug keywords in business names or descriptions. Punctuation and signs were often used to separate each keyword, like "Generation V Vape + CBD Shop | Vapor Shop" and " Buy OG kush online | sour diesel | purple kush | white widow | granddaddy purple." Second, the listings used a set of drugs combining with different promotional terms, such as "Buy Research Chemicals / Order JWH 018 Online, Where to Buy MDPPP Online" in the metadata. Promotional terms are the common terms in the listings' metadata associated with illegal online drug promotion [59]. We found 14,738 (45.32%) of the IDLLs featured promotional terms, of which 3,109 and 14,033 had terms in business names and other metadata, respectively.

V. SECURITY RISKS IN LOCAL LISTING ECOSYSTEM

In this section, we report our investigation study on the security risks of upstream data brokers in the listing ecosystem, including the local listing agents and local data brokers.

A. Weak Vetting of Local Listing Agents

To understand the listing ecosystem from the local listing agents' view, we investigated the services from two popular agents, BrightLocal and MozLocal (see Table V), which are among the most popular listing agents for local SEO [53], and are the representatives for manual and automatic listing data submission/management, respectively. These submission/management channels involve different vetting procedures. Thus, we believe that our investigation results are representative, covering both vetting procedures. This effort aims at understanding the effectiveness and weaknesses of listing data submission from listing agents.

Effectiveness. Both listing agents work with the major local data brokers in the U.S. covering a wide range of local search services such as map (Apple Maps), online rating (Yelp), and social media (Facebook). Customers who hire listing agents need to provide their detailed listings information, including the business name, address, phone number, description, website URL, etc. For the listing data submission, listing agents utilize different approaches that provide insights into how the local listing ecosystem works. Particularly, BrightLocal hires human workers to manually publish listing information to data brokers, yet it does not perform any legality review on the submitted content [15]. MozLocal indicates it automatically submits the listing data to its listing partners [28].

Both submission strategies are effective to get submitted listings published. It will take about one month for BrightLocal to complete the listing process for the submitted listing on its partners, with a cost of about $100/year [15]. Meanwhile, according to MozLocal's customer service, most of its partner brokers will show listings immediately, or it will take up to three weeks to have the listing accessible. The cost of the MozLocal is around $200/year. We also checked how the listing agents update the listing information modifications. Taking MozLocal as an example, based on its service introduction [7], it takes about two weeks to refresh the changes.

Security vetting from listing agents. Based on the communications with the agents and their service introduction [6], [15], the listing agents (BrightLocal and MozLocal) will not vet the listing information during the submission, even for BrightLocal, on which the submission is handled by real people. Through our communication with BrightLocal, which is approved by our institute's IRB, we were told that it does not set any restriction to the content of submitted listing information. The data check of BrightLocal solely focuses on the consistency of the submitted data with the information in business websites and Google My Business profiles, if existing, to ensure that the name, address, and phone number are consistent with the submitted data [15], [16]. Similarly, MozLocal was reported to use a tool to automatically perform data cleansing and validation checks only to ensure that the accurate address format will be accepted by their data partners [6]. From the communication with MozLocal guided by our institute's IRB, the customer service emphasized the requirements of data completeness and format, while no restrictions were requested to the content of the submitted listing. We note that, given the complicated data sharing in the local search ecosystem, a lack of legal and reliability vetting may cause great risks to the downstream parties of being contaminated by polluted listings, which should call for great attention.

Note that we researched the security vetting information of BrightLocal and MozLocal from their websites and customer service communication.[2]

B. Security Risks in Local Data Brokers

Local data sharing. It's a common practice to share data among different brokers in the listing ecosystem, which enriches applications for local intents. However, such cooperation also increases the risks of polluted listing information being propagated in the ecosystem. Indeed, from the communication with the BrightLocal's customer service, we were informed that the listings submitted to some brokers may appear on other sites which are not the partners of BrightLocal because their partner brokers, like Localeze [1], have built a data distribution network with other brokers and sites in the ecosystem. Meanwhile, we were told in the communications—approved by our institute's IRB—with Localeze and BrightLocal that adding listings on both Localeze and BrightLocal has no strict restriction in submitted listings' content (e.g., business names, descriptions). Given the nature of data sharing in the listing ecosystem, it can be expected that polluted listings will appear on many other platforms.

[2]We did not conduct a validation experiment for this paper due to the potential ethical risks to the vetting operators, as suggested by the research ethics committee.

TABLE VI. TOP 10 LOCAL DATA BROKERS AFFECTED BY IDLLS

Data broker	Hit rate*	Data broker	Hit rate
Facebook	63.5%	Merchantcircle	49.7%
Bing	56.6%	Ezlocal	49.0%
Yahoo!	56.3%	Citysearch	47.9%
Insiderpages	54.0%	Centralindex	46.9%
Yellowmoxie	50.1%	Yasabe	45.5%

* The hit rate is calculated by the percentage of the detected IDLLs that can be found from the `Yext.com` query results of each broker.

Validation on local data brokers. From the data-sharing network of listings, we found it is also worrying for the security vetting on the brokers: on many brokers, the submitted listing information is not requested to be verified by these brokers, even when they have content restriction policies [8], [25], [31]. For example, when registering a listing on Facebook, after providing the business name and category, you will be instantly directed to a newly published business page on which you can add the photos, websites, business description, etc. [17]. The verification of the business pages is also optional on Facebook [18]. Similarly, on Foursquare, the submitted listings will be published immediately when the business name, category, and address are provided [19]. Additional evidence has been noted by BrightLocal that some brokers would publish an instant live listing when it submits data to them [15]. Given the data propagation flow from agents to brokers, the polluted listings on agents may be propagated to many downstream parties and even the whole ecosystem without proper vetting from both agents and brokers.

Unclaimed hanging listings. Despite the fact that many local data brokers do not make much effort to vet the listing information published on them, we noticed that even for the brokers which request information verification, business owners can still create listings and leave them unverified while also being available to be reached by users from local search services. For example, it's common to observe unclaimed listings in the search results on Google Maps. For all the IDLLs records discovered in our study, we checked the IDLLs' listing webpages using keywords: "unclaim," "claim your business," "own this business?" etc. We estimated that 61.4% of the IDLLs were unclaimed.

C. Local Business Search Source Credibility

Spread among the data brokers. To understand how IDLLs spread among the data brokers, we define the hit rate, which is the percentage of the detected IDLLs that can be found from the query results of each broker, as returned by `Yext.com`. Note that as mentioned in Section II, `Yext.com`, which aggregates listing information from its 51 partner brokers (e.g., Google Maps, Yahoo!), is the industry's largest listing scan system [37]. In our study, we searched `Yext.com` using the name, address, and phone number of the detected IDLLs, which checked various local brokers to find out whether the listings appeared in the brokers' data collections (e.g., Facebook). Given that 95.88% (see Section III-C) of the sampled IDLLs searched from `Yext.com` can be visible in partner

TABLE VII. TOP 10 POLLUTED UPSTREAM BROKERS ON BING MAPS

Upstream platforms	Num. of IDLLs	Upstream platforms	Num. of IDLLs
Facebook	301	Weedmaps	14
Foursquare	248	Tripadvisor	8
MapQuest	57	Chamber of Commerce	6
Yahoo!	20	Manta	6
Leafly	15	Bizapedia	5

brokers, the discovery of IDLLs in the results of `Yext.com` strongly indicates that the brokers were affected by the IDLLs.

For the 3,571 detected IDLLs, we found 2,549 (77.1%) of them appeared on more than one broker, based on the scanning results from `Yext.com`. Table VI shows the top brokers being affected by these illicit listings. Facebook is the most affected broker with 63.5% of the detected IDLLs affecting it, followed by Bing (56.6%) and Yahoo! (56.3%). More than 55.0% of IDLLs were propagated among at least five brokers, indicating the prevalence of these detected IDLLs in the ecosystem.

In addition to the presence information, we also found that, among the extended listings, 36.2%, 23.3%, and 5.4% of them have different values against the detected IDLLs regarding the name, address, phone number information, and 85.7%, 76.2%, and 45.3% of illicit listings have more than one business name, address, or phone number, respectively. This shows that adversaries always change the listings' names to advertise many different drugs. Regarding the address, 84.5% of the addresses are within the same city of the input detected IDLLs, while the rest are across the U.S. We noticed that a miscreant tends to stick to a small set of phone numbers but use a large number of addresses, names and other identity-related data in their illicit listings, since the phone numbers are the key contact information for buyers.

Propagation to search engines. Although some local data brokers (e.g., Bing My Places) may have a strict vetting process for the listings submitted to them [3], it is interesting that different data vetting policies and efforts enforced by different parties can make the overall vetting less effective.

To estimate IDLL contamination via data propagation in the search engines, we need the data sources of the IDLLs. However, among the search services, only Bing Maps indicates the upstream data sources via a "Data from" section in the search results. We found that among 1,076 IDLLs on Bing Maps, 457 had indicated their upstream data sources in which 169 listings had more than one upstream data source. In total, we identified 37 upstream data sources for those IDLLs on Bing Maps, with Facebook, Foursquare, and MapQuest serving as the top three upstream data sources, which covered 41.45% of the IDLLs on Bing Maps. The top 10 upstream platforms that contaminated Bing Maps are listed in Table VII. In addition to the popular listing platforms, Bing Maps has also been polluted by some professional or relatively unpopular upstream platforms, such as "Weedmaps," "NearSay," "Superpages."

VI. IMPACTS OF ILLICIT DRUG LOCAL LISTINGS

In this section, we studied how the local search results were polluted by IDLLs, which may cause great health concerns to local search end users. To study the impacts of the search results poisoning to the users of voice, web, and map searches, we generated sets of popular queries and checked how the search results were polluted by the detected IDLLs (3,571). As mentioned in Section III-B, we used different keyword resources (e.g., autocompletes, related questions from Google) to generate popular searches, which we call longtail queries. In addition, we used the drug keywords in IDLLs as keyword queries. Each query will be searched at different locations. From the seed keywords of the promoted drugs in the IDLLs, 3,061 keyword queries with different locations and 5,485 longtail queries with different locations were generated for the study. Table VIII lists the number of the IDLLs we discovered using different local search queries with location information on different search channels (Voice, Web, and Map).

A. IDLL Pollution on Voice Search

Here we analyzed the voice search results polluted by the IDLLs on the most popular mobile VA systems, Apple Siri and Google Assistant. As mentioned in Section III-B, speech recognition and OCR were used to trigger and analyze the voice search results. We checked the effectiveness of the speech recognition on these VA systems and found the accuracy to be 92.3% and 91.2% on Apple Siri and Google Assistant, respectively. In the OCR process, we got a 100% accuracy rate for the word recognition.

As mentioned earlier, our idea is to compare the 3,571 detected IDLLs with the results of voice queries. A challenge here, however, is that unlike the listings returned by map and web queries, which are structured, those from voice queries are unstructured, in various formats. Also, the returns from Google Assistant only contain names of the listings while those from Siri include both names and addresses. To find the detected IDLLs from these query results, we utilized the Jaro Similarity [20] between the names (for Google Assistant) and the name-address pairs (for Apple Siri) of each known IDLL and every query result: an IDLL is considered to be identified from the result when the similarity is no less than 0.95. To validate the veracity of this approach and the selected threshold, we randomly sampled 20 identified IDLLs from the search results of each VA system for manual inspection, and found the precision to be 100%.

Voice search polluted by IDLLs. The appearances of listings in the voice search results can be in knowledge panels and map listings. As shown in Table VIII, we found 13.11% of the detected IDLLs can be found on Apple Siri and 25.40% of them are indexed by Google Assistant, indicating the local knowledge databases have been polluted by IDLLs. In our study, Siri only returned one result for each query, and Google Assistant returned five results per query. So, the IDLLs showed in Google Assistant will be among the top five results, and those reported by Siri were always the first ones.

When looking into the voice commands' pollution rate, 11.09% and 14.33% of the voice search queries were polluted by IDLLs on Apple Siri and Google Assistant in Table VIII, which could indicate more credible data sources used by Apple Siri than that of Google Assistant.

We also observed the polluted voice queries are the common questions asked by users (see Section III-B), with an average length of five (without location terms). Such queries always have large search volumes (see Section VI-C) and

therefore will potentially affect the increasingly growing voice search audience. Interestingly, we noticed that, among the detected IDLLs found from voice search, 14.15% of them can be encountered by more than one voice command. For example, seven voice queries (e.g., "where to buy bath salt," "how to make research chemical acid") were polluted by the IDLL "Legal Bath Salts and Research Chemicals Online Shop" when searching in Georgia. This is because the IDLLs usually feature several drug keywords in their profile. Thus, they will be searched with different queries.

B. IDLL Impacts on Web and Map Search

We further studied the local search poisoning of the IDLLs in the web's knowledge panels and the map's listings on five web and map search engines, including Apple Maps, Bing Web and Maps, and Google Web and Maps. For map search, we collected the search results on the first page.

Local business search pollution rate. Table VIII shows that IDLLs are more likely to be encountered through map search than web search, which will cause real harm to the local search users who usually search on maps for local intents. More specifically, 12.27% of the detected IDLLs can be encountered by the queries on Bing Maps, which is 11.87% and 7.98% on Google Maps and Apple Maps, respectively. We looked into the total number of queries that can lead to IDLLs on each map search platform (among the total 8,546 queries made on the platform, see Table VIII). As we can see, queries for local map search on Google are most likely to introduce IDLLs, with 1,599 (18.71%) of them leading to IDLLs, followed by 1,114 (13.04%) on Bing and 302 (3.53%) on Apple.

C. IDLL Impacts on Search Environments

IDLL prevalence between different channels. Table VIII shows the IDLL prevalence in voice, web, and map searches. More IDLLs can be found from voice search than web or map search in Apple and Google. Also, more queries can lead to IDLLs through voice or map search than web search. These findings indicate that voice search, an emerging search channel, likely suffers more from IDLLs than traditional web or map search, though map search is also seriously affected compared with web search. Followed by Apple and Bing, Google is the most affected search engine, where more IDLLs have been found, and more queries can lead to them.

Thus, immediate actions should be taken to mitigate the IDLL pollution on local search results, which would cause serious concerns about their trustworthiness. Note that the IDLLs should not display in the knowledge panels, maps, and voice search, even to the users who are actively seeking drug-related results, according to the search engines' policy [13]. Also, we found that illicit listings would show up even when the queries are benign, like "420" and "bath salts."

IDLLs search volumes and revenue. We studied the search volumes of the queries that can lead to IDLLs, leveraging `KeywordTool.io` [21] and `Wordtracker.com` [36] to get the average monthly search volumes of these queries in the past 12 months. The average search volumes of keyword queries and longtail queries are 137,751 and 1,090, respectively. A high search volume demonstrates that the IDLLs introduced by our queries would be exposed to a large portion of the search users, posing a huge threat to the individuals and society. According to the finding in [62], the web search conversion rate for online drug sales is between 0.3% to 3%. For example, the monthly search volume of the longtail keyword "buy research chemicals online" is 517. From the homepage of one introduced IDLL (`www.omegachemicalsonline.com`), the cheapest product "5F-AKB48 (10g)" is priced $200 per item, whose initial dependence usage is between 1˜4 grams/day and the dose increases quickly afterwards [71]. Assuming the conversion rate is 1% and each buyer consumed 2 grams/day, the monthly revenue from new buyers would be $6,204. Every year the store would accumulate more than 400 buyers. Even if they buy one package of the cheapest product each month, the monthly revenue would be more than $480,000. With the accumulation of the buyers and increasing demand from addicted buyers, the revenue would be much more considerable.

TABLE VIII. IDLLS DISCOVERED BY SEARCH QUERIES ON DIFFERENT SEARCH CHANNELS

Search channels	Apple #IDLL|#Query*	Bing #IDLL|#Query	Google #IDLL|#Query
Voice	468 | 948	-	907 | 1,225
Web	-	85 | 157	60 | 104
Map	285 | 302	438 | 1,114	424 | 1,599
Total	677 | 1,242	460 | 1,187	1,206 | 2,808

* `#IDLL` is the number of the detected IDLLs on the search engine. `#Query` is the number of the queries that can lead to IDLLs on the search engine.

VII. DISCUSSION

In this section, we propose some intervention strategies to mitigate this emerging threat. The local business ecosystems outside the U.S. are further considered and compared. Meanwhile, we discuss method limitations and our legal and ethical considerations.

Intervention strategies. Based on our understanding of the local search poisoning for online illicit drug promotions, we propose several intervention strategies to mitigate this emerging and underestimated threat. Specifically, **(1)** as the key data source in the local search ecosystem, we suggest the local data brokers and listing agents should take responsibility in conducting more strict security vetting for the listing information published on them by individual owners or their "trusted data contributors." **(2)** Stakeholders in the local search ecosystem should unify their listing data quality policies to disrupt the propagation of polluted data from other data sources. **(3)** Local search services should make more of an effort to vet the shared data from upstream brokers as if they were provided by any untrusted party. They and escrow supervisors should take more actions to delist those poisoned local search results. Similar to the efforts of the Safe Browsing service, local search powered by local knowledge bases should be equipped with such security protection. We propose a new technique, IDLLSpread, for detecting IDLLs, which can be directly used by data brokers, listing agents, and local search services.

Ecosystems outside the U.S. Except for the measurement of the local business ecosystem in the U.S., we also looked into the ecosystems in the U.K., Canada, Brazil, and Germany [4],

[5], [10], [32]. Compared with the U.S. ecosystem, those in non-U.S. countries have some similarities and differences. The first similarity is that all these ecosystems have data flow from data brokers to search engines. Second, many brokers (e.g., Foursquare, Factual, Yelp) and search engines (Google, Bing, Apple) still play the same roles in these ecosystems as they do in the U.S. The third is that the listing agents, like BrightLocal and MozLocal, can also be hired in some countries such as Germany, the U.K., and Canada for listing publication [28]. However, two main differences exist in different ecosystems. First, in countries like Canada and the U.K., government entities offer local listing databases to the public and this data also flows into the brokers and search engines [5], [32]. Second, some data brokers are only active in specific countries, such as Scoot in the U.K. and Gelbe Seiten in Germany.

Limitations. Our IDLLs detection may be biased, given that labeled illicit listings were gathered using heuristics of promotional signals. An IDLL can be disguised as a normal listing without advertising its drug products, which may not be detected by IDLLSpread. However, such behaviors will also make the IDLLs less visible to users and the promotion less successful. Whether or not the listings and their websites were actually selling drugs could only be known after we made a payment and received the drug package. Given the ethical concerns, we mostly focused on the intention (i.e., illicit drug promotion) of the listings from their names, websites, descriptions, etc., and analyzed those that promote drugs. Given the policies and guidance of some illegal drugs (e.g., marijuana, kratom) vary between the federal and state levels, state and local levels, or among different states, it is very difficult if not impossible to judge whether some listings are, indeed, illegal. In our study, we made our best effort to reduce the false positives by applying simple yet effective filtering rules on IDLLSpread's detection results. It is well-known that drug legislation is complicated, and we encourage the legislation department, drug enforcement administration, and listing ecosystem to work together to regulate the drug listings.

Legal and ethical considerations. In our study, we communicated with the customer services of BrightLocal, MozLocal, Localeze, etc. to understand their listing vetting and data-sharing policies. This experiment had been determined and approved by our institute's IRB as "Not Human Subjects Research." We worked with our IRB counsel to design the communications with customer services of the agents (i.e., BrightLocal and MozLocal) and brokers (e.g., Localeze) via email to ensure that we acted under a legal and ethical framework that minimized any risk of harm to any party. In addition, we responsibly disclosed our findings to the affected data brokers and search engines like Infogroup, Google, Bing, and Apple [23]. So far, we have not received any response.

VIII. Related work

Detecting search poisoning. SEO campaigns have been extensively studied [63], [64], [69]. Particularly, the prior research [73] measured the SEO campaigns from web search results and focused on crawled website content to extract features and analyze the campaign. From a different perspective, our paper focuses on the listings reported by local knowledge panels, map search, and voice search. Compared with web search results, listings on these sources are characterized by limited information (only name, phone number, address, URL, description, category), and therefore make it much harder to detect the poisoned content they carry. To address this challenge, IDLLSpread leverages the observation that illicit listings tend to be related so we can utilize the metadata (name, address, description, etc.) shared across different listings for detection. Mining on the graph built on top of these relations is shown to be effective in capturing IDLLs.

Detecting local listing spam. To our best knowledge, the prior work on measuring spam listings on Google Maps [56] is the most relevant one to our study. From the search service insider's view, prior research analyzed the abusive user-generated listings on the map, which were detected by Google's scanning algorithm. Unlike prior work, which just focused on understanding individual spam cases, our approach (IDLLSpread) aims at uncovering the underlying relations among illicit pharmacy listings, which are shown to be an effective means for detecting new illicit cases. For this purpose, we ran graph mining on the dataset collected from local search services' upstream data brokers and utilized the connectivity and clustering of discovered listings to identify new ones. We also measured the impact of such poisoning attacks on prominent local search services. This effort leads to the identification of SEO campaigns, not just single cases as reported in the prior work [56]. However, there is no public data available, up to our knowledge, about Bing and Apple's effort in removing pollution listings. Nor do we see any indicator about the listings of these search engines later being removed. Also importantly, the inherent relations revealed by our approach enable the study of the whole ecosystem behind illicit drug listings and their strategies (e.g., using brokers instead of direct listings on search engines). Even on the level of individual cases, we brought to light a set of new discoveries, such as impersonating similar but legitimate local businesses to promote online stores (see Section IV-C).

Study on illicit online drug promotion. The abuse of addictive and non-prescription medications is a national epidemic. Many studies have been done to identify illegal promotion of such controlled substances on the Internet, such as online marketplaces, search engines, and social media. Specifically, the prior studies [55], [70] measured the online anonymous marketplace Silk Road and revealed that sixteen of the top twenty products sold were illicit, drug-related controlled substances. Leontiadis, etc. [62] focused on the analysis of search-redirection attacks for illicit online drug promotion on the search engine in which websites were compromised to redirect users' search traffics to online pharmacies. Researchers had adopted unsupervised machine learning and data mining to identify online drug promotion: the prior research [58], [59], [66] utilized Biterm Topic Model to identify promotional topics in Twitter and characterized different types of illegal online drug marketing. Also, deep learning models were used to detect illicit drug dealing on Instagram on which the drugs were promoted via sharing images and videos [74].

In contrast, our work is the first to provide a detailed analysis of illicit online drug promotion through local search poisoning, which is an emerging promotion channel targeting the increasing local search intent. Our work also brings

invaluable insights into the security risks in the local listing ecosystem and the broad impacts of the poisoned listings.

IX. Conclusion

In this paper, we report the first systematic study of IDLLs on local search services in which the miscreants pollute the search results returned from local search services (e.g., local knowledge panel, map search, and voice search) for illicit drug promotion by posting promotional listings on the upstream local data brokers with loose control on data quality. In our research, we proposed IDLLSpread, a new methodology for IDLL discovery and tracking, which utilizes semi-supervised graph mining to discover IDLLs from the upstream local data brokers and further analyzes their reachability to the downstream local search service audience. Running on the collected 94,856 listings, IDLLSpread reported 3,571 IDLLs, which can be reached by 2.73%, 30.68%, and 25.20% of drug-related queries through local knowledge panels, map search, and voice search, respectively. After the extension, we returned 32,520 IDLLs in total. Our study sheds light on the scope, impact, and techniques of such trending illicit promotional activities. We further investigated the security protection of today's local listing ecosystem, which reveals that the upstream data providers are less regulated and lack proper vetting procedures, allowing the spread of contaminated information to prominent downstream listing services. Our findings and techniques contribute to a better understanding of today's local search ecosystem and enhancement of protection for search audience.

Acknowledgment

We would like to thank the anonymous reviewers for their insightful comments. This work is supported in part by the NSF CNS-1801432, 1838083, and 1850725.

References

[1] "About localeze," https://www.neustarlocaleze.biz/app/help/about-localeze.

[2] "Amazon polly," https://aws.amazon.com/polly/.

[3] "Bing places for business - verification options," https://www.bingplaces.com/Help#VerificationOptions.

[4] "Brazilian local search ecosystem," https://moz.com/learn/seo/local-search-data-brazil.

[5] "Canadian local search ecosystem," https://moz.com/learn/seo/local-search-data-canada.

[6] "Data cleansing - help hub - moz," https://moz.com/help/moz-local/manage-locations/data-cleansing.

[7] "Directories - help hub - moz," https://moz.com/help/moz-local/manage-locations/directories.

[8] "Foursquare consumer services terms of use," https://foursquare.com/legal/terms.

[9] "Generating related questions for search queries," patents.google.com/patent/US9213748B1/en.

[10] "German local search ecosystem," https://moz.com/learn/seo/local-search-data-europe.

[11] "Google autocomplete," https://blog.google/products/search/how-google-autocomplete-works-search/.

[12] "Google business profile," https://support.google.com/business/answer/6331288?hl=en.

[13] "Google maps prohibited and restricted content," https://support.google.com/contributionpolicy/answer/7400114?hl=en.

[14] "Google ocr," cloud.google.com/vision/docs/ocr.

[15] "How is a citation builder campaign delivered? – brightlocal help center," https://help.brightlocal.com/hc/en-us/articles/360033890793-How-is-a-Citation-Builder-campaign-delivered-.

[16] "How long does it take for my listings to be updated through data aggregator submissions using citation builder? – brightlocal help center," https://help.brightlocal.com/hc/en-us/articles/360036410794-How-long-does-it-take-for-my-listings-to-be-updated-through-Data-Aggregator-submissions-using-Citation-Builder-.

[17] "How to create a facebook business page in 5 steps," https://fitsmallbusiness.com/how-to-create-a-facebook-business-page.

[18] "How to get your facebook page verified," https://www.adquadrant.com/blog/how-to-get-your-facebook-page-verified.

[19] "How to submit a business listing to foursquare," https://www.advicelocal.com/blog/how-to-submit-a-business-listing-to-foursquare.

[20] "Jaro–winkler distance - wikipedia," https://en.wikipedia.org/wiki/Jaro-Winkler_distance.

[21] "Keyword tool api," https://keywordtool.io/api.

[22] "Kratom legality," https://kraoma.com/kratom-legality-united-states/.

[23] "Local listing poisoning," https://sites.google.com/view/idlls.

[24] "Local search ecosystem," https://https://whitespark.ca/local-search-ecosystem.

[25] "Manta terms of service," https://www.manta.com/resources/page_terms_conditions.

[26] "Marijuana retail stores with medical endorsement list," https://www.doh.wa.gov/Portals/1/Documents/Pubs/608017.pdf.

[27] "Mmtc - ommu," https://knowthefactsmmj.com/mmtc/.

[28] "Moz local vs brightlocal — compare moz local alternatives," https://www.brightlocal.com/moz-local-vs-brightlocal/.

[29] "Oregon liquor control commission: Active marijuana retail licenses approved as of 1/15/2021," https://www.oregon.gov/olcc/marijuana/Documents/Approved_Retail_Licenses.pdf.

[30] "Oregon revised statutes (ors) 475b: Cannabis regulation," https://www.oregonlegislature.gov/bills_laws/ors/ors475B.html.

[31] "Terms and policies — facebook," https://www.facebook.com/policies_center/commerce.

[32] "Uk local search ecosystem," https://moz.com/learn/seo/local-search-data-uk.

[33] "Verify your local business on google - google my business help," https://support.google.com/business/answer/7107242.

[34] "Voice search stats," www.dialogtech.com/blog/voice-search-statistics/.

[35] "Wayback machine scraper," https://github.com/sangaline/wayback-machine-scraper.

[36] "Wordtracker keyword api 2.0," https://www.wordtracker.com/api.

[37] "Yext listings overview," https://help.yext.com/hc/en-us/articles/360001288866-Listings-Overview.

[38] "Yext listings scan," https://www.yext.com/pl/powerlistings/scan.html.

[39] "Ryan haight act," www.justice.gov/archive/olp/pdf/hr-6353-enrolled-bill.pdf, 2008.

[40] "Google fined $500m for illegal drug ads," www.justice.gov/opa/pr/google-forfeits-500-million-generated-online-ads-prescription-drug-sales-canadian-online, 2011.

[41] "Bing knowledge," https://blogs.bing.com/search-quality-insights/2017-07/bring-rich-knowledge-of-people-places-things-and-local-businesses-to-your-apps, 2017.

[42] "What are data aggregators?" https://www.advicelocal.com/blog/data-aggregators-local-business/, 2017.

[43] "Dea drug slang term and code words," https://www.dea.gov/documents/2018/07/01/2018-slang-terms-and-code-words, 2018.

[44] "2019 voice report," https://about.ads.microsoft.com/en-us/insights/2019-voice-report, 2019.

[45] "Commonly used drugs charts," https://www.drugabuse.gov/drug-topics/commonly-used-drugs-charts/, 2020.

[46] "Data aggregators submissions," https://www.brightlocal.com/citation-builder/local-data-aggregators/, 2020.

[47] "Marijuana guide," https://potguide.com/, 2020.

[48] "Opioid treatment directory," https://dpt2.samhsa.gov/treatment/, 2020.

[49] "Samhsa treatment facilities," www.samhsa.gov/data/report/national-directory-drug-and-alcohol-abuse-treatment-facilities-2020, 2020.

[50] "Textoptimizer," https://textoptimizer.com, 2020.

[51] "Top 50 local citation sites for the u.s," https://www.brightlocal.com/resources/top-50-local-citation-sites/, 2020.

[52] "Top citation sites in the u.s," https://whitespark.ca/top-local-citation-sources-by-country/united-states/, 2020.

[53] "Local seo tools: Top 10 tools to improve your local search ranking," https://marketingdisty.com/10-best-local-seo-tools/, 2021.

[54] "Networkx," https://networkx.org/, 2021.

[55] N. Christin, "Traveling the silk road: A measurement analysis of a large anonymous online marketplace," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 213–224.

[56] D. Y. Huang, D. Grundman, K. Thomas, A. Kumar, E. Bursztein, K. Levchenko, and A. C. Snoeren, "Pinning down abuse on google maps," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 1471–1479.

[57] J. Isacenkova, O. Thonnard, A. Costin, A. Francillon, and D. Balzarotti, "Inside the scam jungle: A closer look at 419 scam email operations," *EURASIP Journal on Information Security*, vol. 2014, no. 1, pp. 1–18, 2014.

[58] J. Kalyanam, T. Katsuki, G. R. Lanckriet, and T. K. Mackey, "Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the twittersphere using unsupervised machine learning," *Addictive behaviors*, vol. 65, pp. 289–295, 2017.

[59] J. Kalyanam and T. Mackey, "Detection and characterization of illegal marketing and promotion of prescription drugs on twitter," *arXiv preprint arXiv:1712.00507*, 2017.

[60] T. Katsuki, T. K. Mackey, and R. Cuomo, "Establishing a link between prescription drug abuse and illicit online pharmacies: analysis of twitter data," *Journal of medical Internet research*, 2015.

[61] R. Kyng, A. Rao, S. Sachdeva, and D. A. Spielman, "Algorithms for lipschitz learning on graphs," in *Conference on Learning Theory*, 2015, pp. 1190–1223.

[62] N. Leontiadis, T. Moore, and N. Christin, "Measuring and analyzing search-redirection attacks in the illicit online prescription drug trade." in *USENIX Security Symposium*, vol. 11, 2011.

[63] ——, "A nearly four-year longitudinal study of search-engine poisoning," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014, pp. 930–941.

[64] X. Liao, C. Liu, D. McCoy, E. Shi, S. Hao, and R. Beyah, "Characterizing long-tail seo spam on cloud web hosting services," in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 321–332.

[65] X. Liao, K. Yuan, X. Wang, Z. Pei, H. Yang, J. Chen, H. Duan, K. Du, E. Alowaisheq, S. Alrwais *et al.*, "Seeking nonsense, looking for trouble: Efficient promotional-infection detection through semantic inconsistency search," in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 707–723.

[66] T. K. Mackey, J. Kalyanam, T. Katsuki, and G. Lanckriet, "Twitter-based detection of illegal online sale of prescription opioid," *American journal of public health*, vol. 107, no. 12, pp. 1910–1915, 2017.

[67] G. Mba, J. Onaolapo, G. Stringhini, and L. Cavallaro, "Flipping 419 cybercrime scams: Targeting the weak and the vulnerable," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 1301–1310.

[68] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[69] A. Oest, Y. Safaei, A. Doupé, G.-J. Ahn, B. Wardman, and K. Tyers, "Phishfarm: A scalable framework for measuring the effectiveness of evasion techniques against browser phishing blacklists," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 1344–1361.

[70] K. Soska and N. Christin, "Measuring the longitudinal evolution of the online anonymous marketplace ecosystem," in *24th USENIX security symposium*, 2015, pp. 33–48.

[71] M. C. Van Hout and E. Hearne, "User experiences of development of dependence on the synthetic cannabinoids, 5f-akb48 and 5f-pb-22, and subsequent withdrawal syndromes," *International Journal of Mental Health and Addiction*, vol. 15, no. 3, pp. 565–579, 2017.

[72] B. Wang, N. Z. Gong, and H. Fu, "Gang: Detecting fraudulent users in online social networks via guilt-by-association on directed graphs," in *2017 IEEE International Conference on Data Mining (ICDM)*.

[73] D. Y. Wang, M. Der, M. Karami, L. Saul, D. McCoy, S. Savage, and G. M. Voelker, "Search + seizure: The effectiveness of interventions on seo campaigns," in *Proceedings of the 2014 Conference on Internet Measurement Conference*, 2014, pp. 359–372.

[74] X. Yang and J. Luo, "Tracking illicit drug dealing and abuse on instagram using multimodal analysis," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 4, pp. 1–15, 2017.

[75] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in neural information processing systems*, 2004, pp. 321–328.

[76] W. Zhu, H. Gong, R. Bansal, Z. Weinberg, N. Christin, G. Fanti, and S. Bhat, "Self-supervised euphemism detection and identification for content moderation," in *IEEE Symposium on Security and Privacy (SP). IEEE*, 2021.