

Measuring and Analyzing Search Engine Poisoning of Linguistic Collisions

Matthew Joslin*, Neng Li[†], Shuang Hao*, Minhui Xue[‡], Haojin Zhu [†]

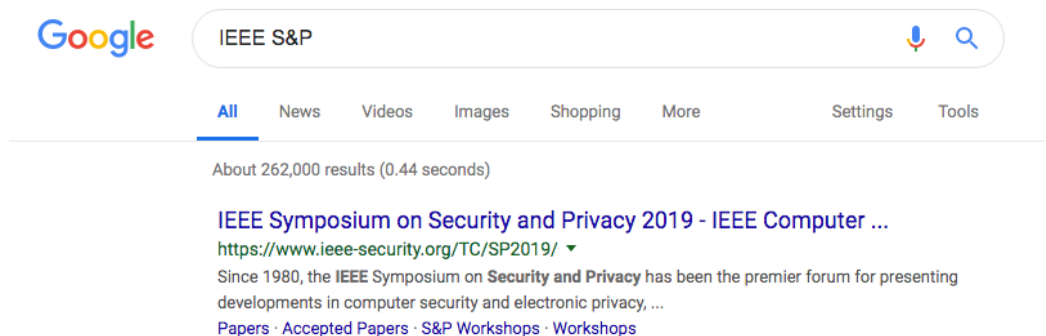
*University of Texas at Dallas, [†] Shanghai Jiao Tong University,

[‡] Macquarie University

{matthew.joslin, shao}@utdallas.edu {ln-fjpt, zhu-hj}@sjtu.edu.cn minhuixue@gmail.com



Search Rank Dominates Web Traffic



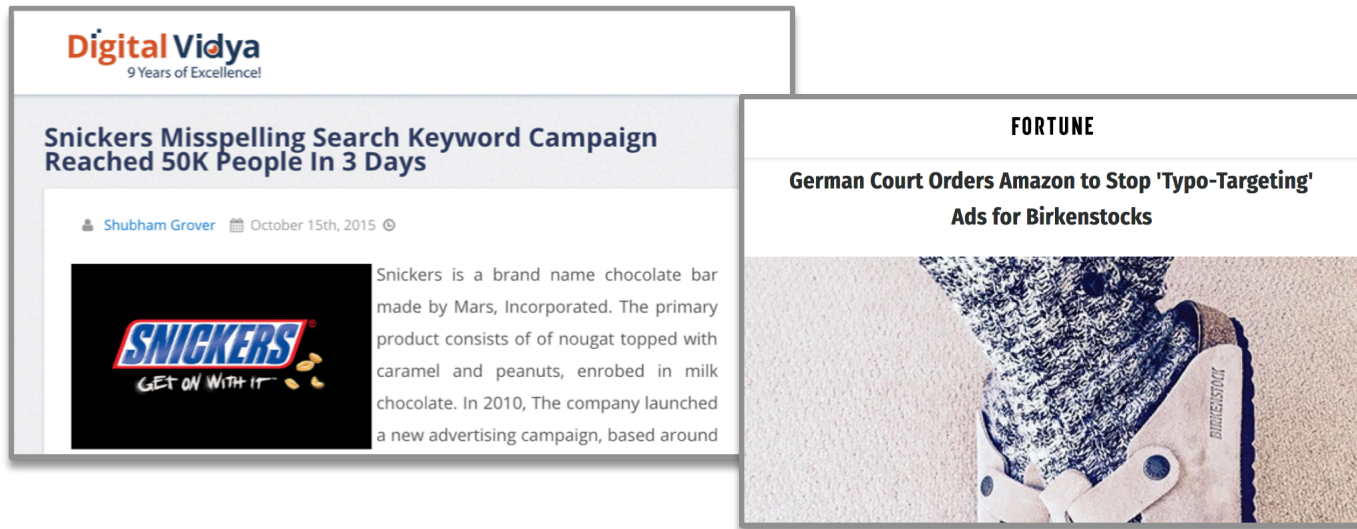
- ▶ **51%** of traffic from web search
- ▶ **90%** of users click search results returned on the first page

Source: *Search Engine Land* and *ProtoFuse*

Google and the Google logo are registered trademarks of Google LLC, used with permission.

Searches with Misspelled Keywords

- Users make mistakes when typing searches
 - **adoeb** (a misspelling of **adobe**)



Auto-Correction and Auto-Suggestion

Misspelling

adoeb

Showing results for [adobe](#)
Search instead for [adoeb](#)

[Adobe: Creative, marketing and document management solutions](#)
[www.adobe.com/](#) ▼
Adobe is changing the world through digital experiences. We help our customers create, deliver and optimize content and applications.

[Adobe \(@Adobe\)](#) [Twitter](#)
[https://twitter.com/Adobe](#)

adoeb

Showing results for ...

- High confidence

Misspelling

adobec

Including results for [adobe](#)
Search only for [adobec](#)

[Adobe: Creative, marketing and document management solutions](#)
[www.adobe.com/](#) ▼
Adobe is changing the world through digital experiences. We help our customers create, deliver and optimize content and applications.

[adobec - Roblox](#)
[https://www.roblox.com/users/92989488/profile](#) ▼
adobec is one of millions playing, creating and exploring the endless possibilities of Roblox. Join adobec on Roblox and explore together!

adobec

Including results for...

- Medium confidence

Misspelling

adube

Did you mean: [adube](#)

[adube - Wiktionary](#)
[https://en.wiktionary.org/wiki/adube](#) ▼
Ido[edit]. Adverb[edit]. adube, where. Retrieved from "https://en.wiktionary.org/w/index.php?title=adube&oldid=46316325". Categories: Ido lemmas · Ido adverbs. Navigation menu. Personal tools. Not logged in; Talk · Contributions · Preferences · Create account · Log in. Namespaces. Entry · Discussion · Citations. Variants ...

[Significado / defini](#) [-ro de adube no Dicion](#) [-rio Priberam da L](#) [-ngua...](#)
[https://www.priberam.pt/dlpo/adube](#) ▼ [Translate this page](#)
Significado / defini [-ro de adube no Dicion](#) [-rio Priberam da L](#) [-ngua Portuguesa.](#)

adube

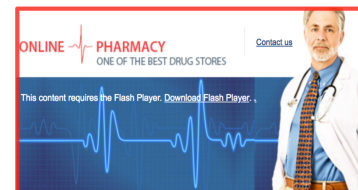
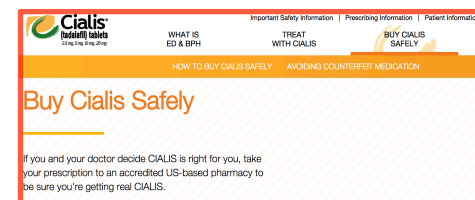
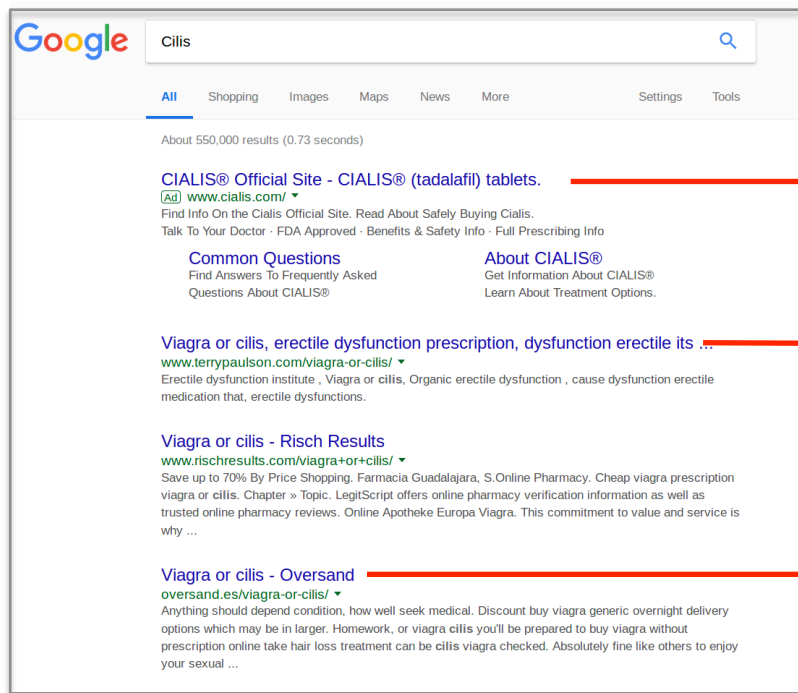
Did you mean...

- Low confidence

Linguistic-Collision Misspellings



Cilis
(misspelling
of **Cialis**)

In Esperanto:
“chilis”



Google and the Google logo are registered trademarks of Google LLC, used with permission.

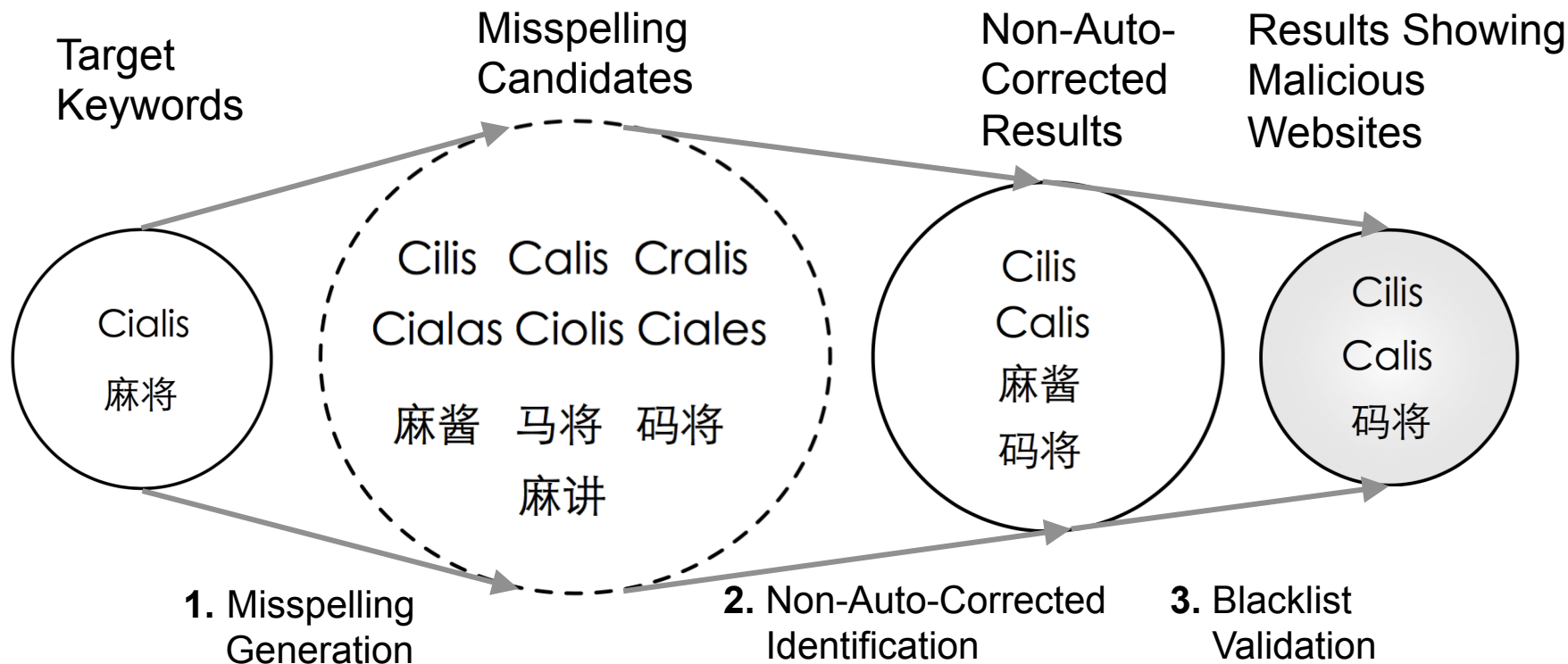
Study Scope

- ▶ Analyzed languages
 - English and Chinese
- ▶ Search engines
 - Google and Baidu  
- ▶ Target keywords
 - Alexa 10k domains (English only)
 - 13 selected categories

Keyword Categories

- ▶ 4 spam-related categories: drugs, adult, gambling, software
 - English examples: Cialis, poker
 - Chinese examples: 大麻, 麻將
- ▶ 9 other categories: cars, food, jewelry, women's clothing, men's clothing, cosmetics, baby products, daily necessities, defense contractors

Our Approach

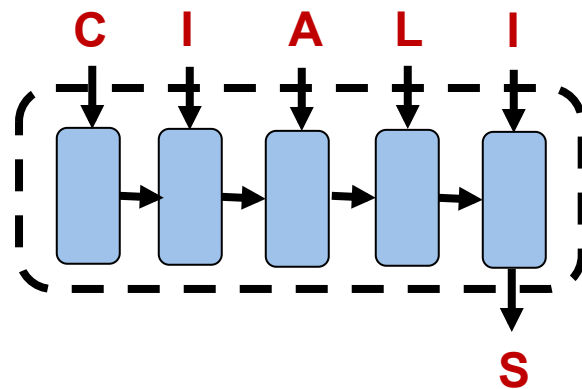


English Misspelling Generation

- ▶ Damerau-Levenshtein edit distance one
 - Insert: cia**l**lis
 - Replace: ci**o**lis (Limited to adjacent keys on QWERTY)
 - Transpose: ca**il**lis
 - Delete: ci**a**lis
- ▶ Vowel replacement
 - a, e, i, o, u, y

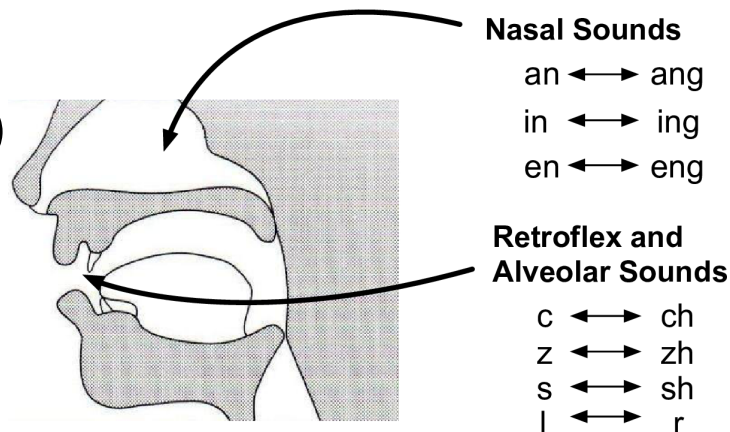
Predicting Linguistic Collision Misspellings

- ▶ Brute-force checking is too time-consuming
- ▶ Dictionaries have poor coverage
- ▶ Using character-level Recurrent Neural Network (RNN) to predict
 - Training with existent words from dictionaries

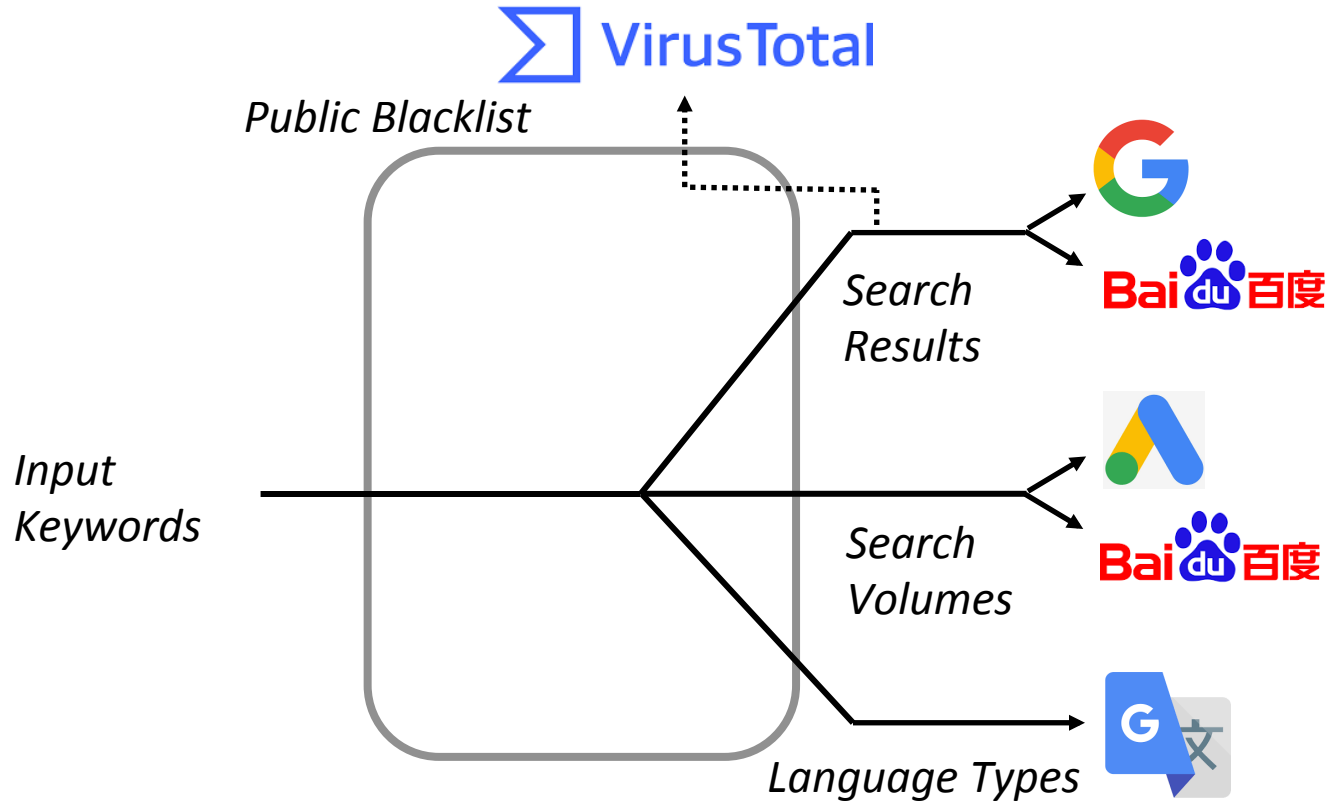


Chinese Misspelling Generation

- ▶ Pinyin input
 - Method for typing Chinese words with the English alphabet
- ▶ Damerau-Levenshtein edit distance one
- ▶ Same pinyin or different tones
 - MáJiàng: 麻將 (tile-based game)
or 麻酱 (sesame sauce)
- ▶ Fuzzy pinyin



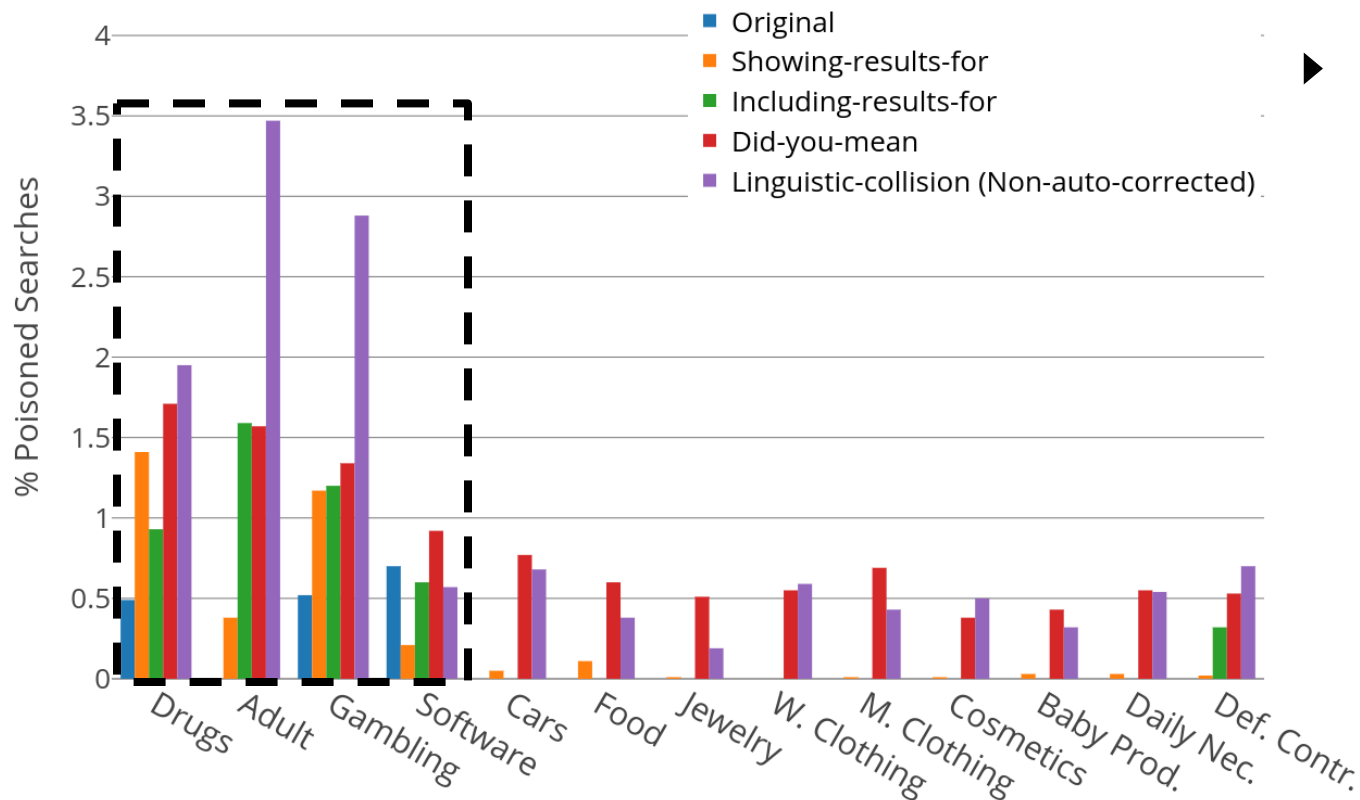
Crawling Framework



Overall Statistics

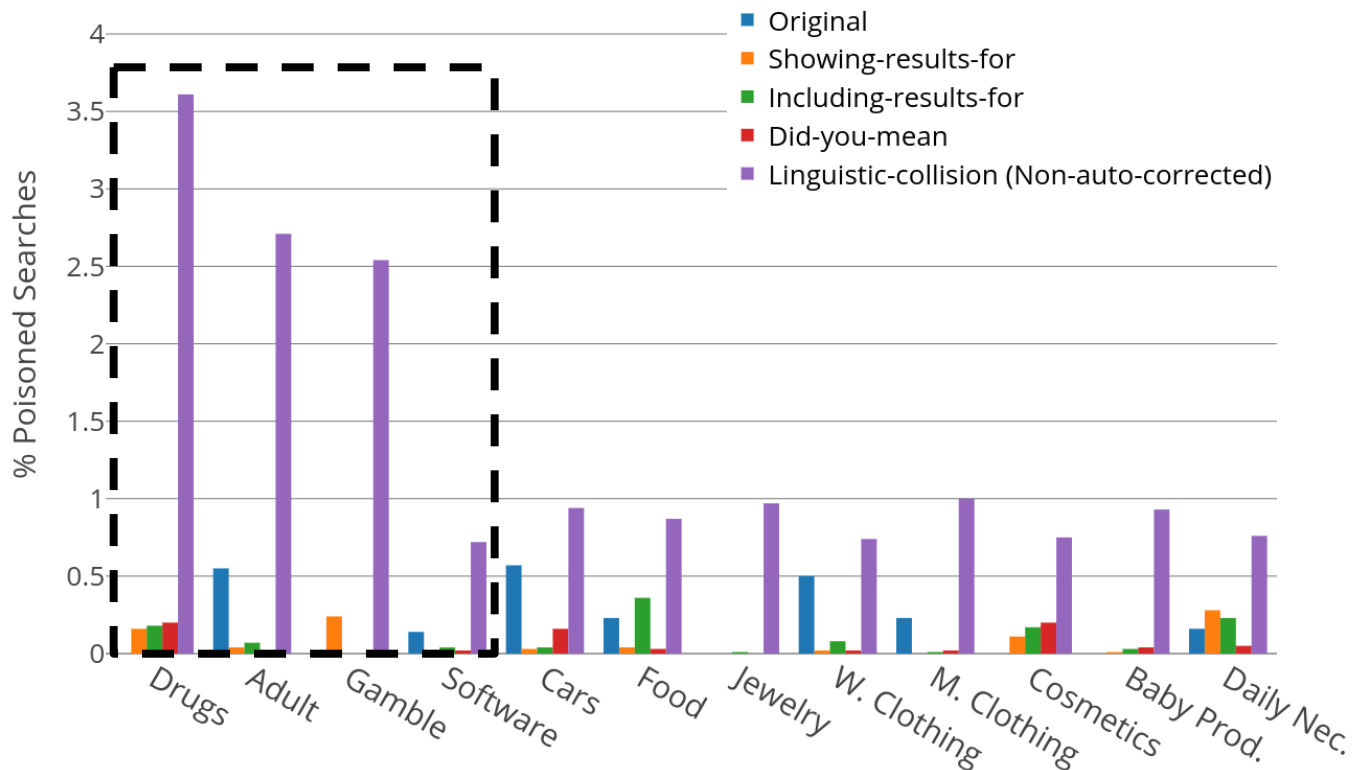
- ▶ 1.77M misspelling candidate keywords queried
- ▶ 1.19% of linguistic-collision misspellings have search results with blacklisted URLs on the first page (10 results per page)

Prevalence: English Search Poisoning



- **Drugs, adult, and gambling** categories targeted at 4x the rate of others

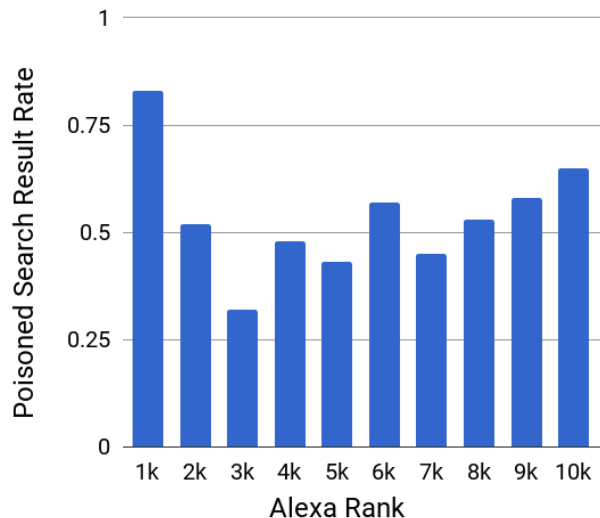
Prevalence: Chinese Search Poisoning



- ▶ Auto-corrected cases exhibit lower poisoning than English.

Results on Alexa List

- ▶ Alexa 1k
 - Exhaustive search to compare with RNN results
 - RNN is 2.84x more efficient than random sampling
- ▶ Alexa 10k
 - Used RNN to generate linguistic collision candidates
 - Attackers exhibit activity across the long tail of domains



Traffic Breakdown per Device Types

	<i>English</i>		<i>Chinese</i>	
<i>Device Type</i>	<i>Original Keywords</i>	<i>Misspellings Targeted by Attackers</i>	<i>Original Keywords</i>	<i>Misspellings Targeted by Attackers</i>
Desktop	36.05%	11.96%	39.74%	21.22%
Mobile	56.56%	84.56%	60.26%	78.78%
Tablet	7.40%	3.48%	----	----

- ▶ English data from Google Adwords
- ▶ Chinese data from Baidu Index

Top English Malicious Domains

<i>Domain Name</i>	<i># of Poisoned Searches</i>	<i># of URLs</i>	<i>Traffic Monetization</i>
*.0catch.com	732	109	malvertising
*.atspace.name	63	17	malvertising
hdvidzpro.me	58	58	malvertising
wanna[REDACTED].com	49	48	malvertising
theunderweardrawer.co.uk	40	38	malvertising

Linguistic Collision Languages

<i>All Results</i>		<i>Drugs</i>		<i>Gambling</i>		<i>Adult Terms</i>	
English	57.44%	English	49.28%	English	66.44%	English	81.67%
Arabic	2.76%	Latin	3.69%	Spanish	2.69%	French	1.96%
Spanish	1.66%	Spanish	2.82%	Norwegian	2.14%	Spanish	1.30%
Hindi	1.56%	Italian	2.47%	Italian	1.78%	Indonesia	1.05%
Italian	1.53%	Romanian	2.25%	French	1.68%	Polish	0.79%

- Languages identified by Google Translate

Conclusion

- ▶ First investigation into linguistic collisions for **English and Chinese**
- ▶ **1.19%** of linguistic-collision misspellings have search results with blacklisted URLs on the first page
- ▶ Certain categories are more heavily targeted and mobile users are more likely to search poisoned terms

Q&A

Thank you!

matthew.joslin@utdallas.edu

Collisions: Statistics

- ▶ Non-auto-corrected:
 - 15.16% English
 - 7.69% Chinese
- ▶ Misspelling methods:
 - Wrong vowel: 22.85% (English)
 - Same pronunciation: 18.21% (Chinese)
 - Fuzzy pinyin: 17.63% (Chinese)