DSPIN: Detecting Automatically Spun Content on the Web

Qing Zhang, David Y. Wang, Geoffrey M. Voelker

University of California, San Diego

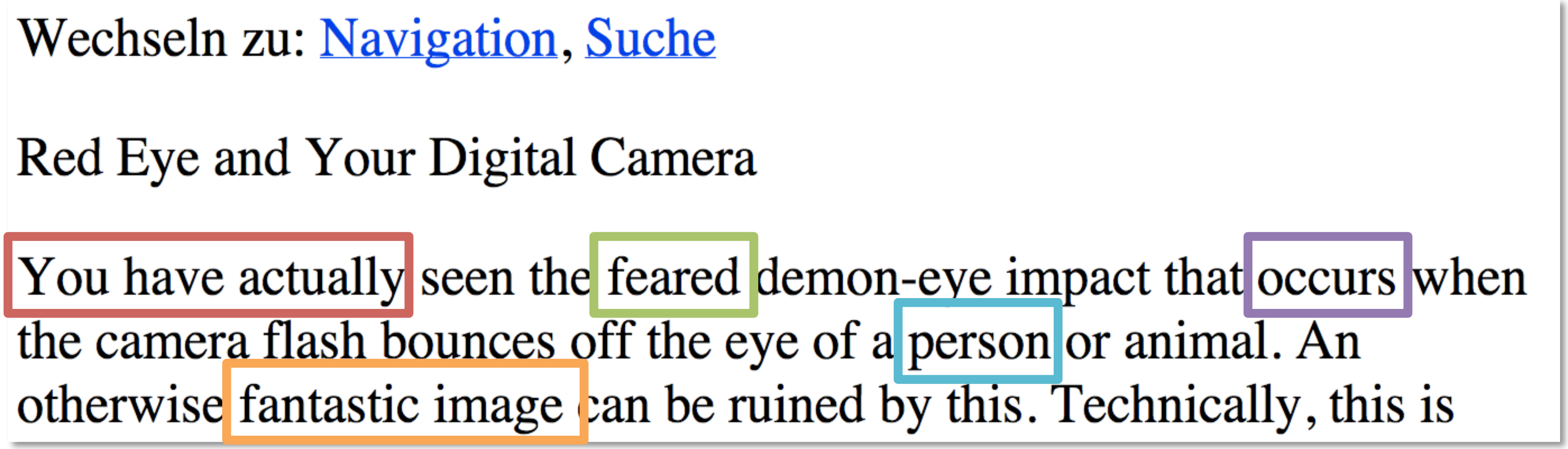
UCSDCSE
Computer Science and Engineering

What is Spinning?

- A **B**lack **H**at **S**earch **E**ngine **O**ptimization **(BHSEO)** technique that rewords original content to avoid duplicate detection
- Typically an article (seed) is spun multiple times creating *N* versions of the article that will be posted on *N* different sites
- Artificially generate interest to increase search result rankings of targeted site

UCSDCSE
Computer Science and Engineering

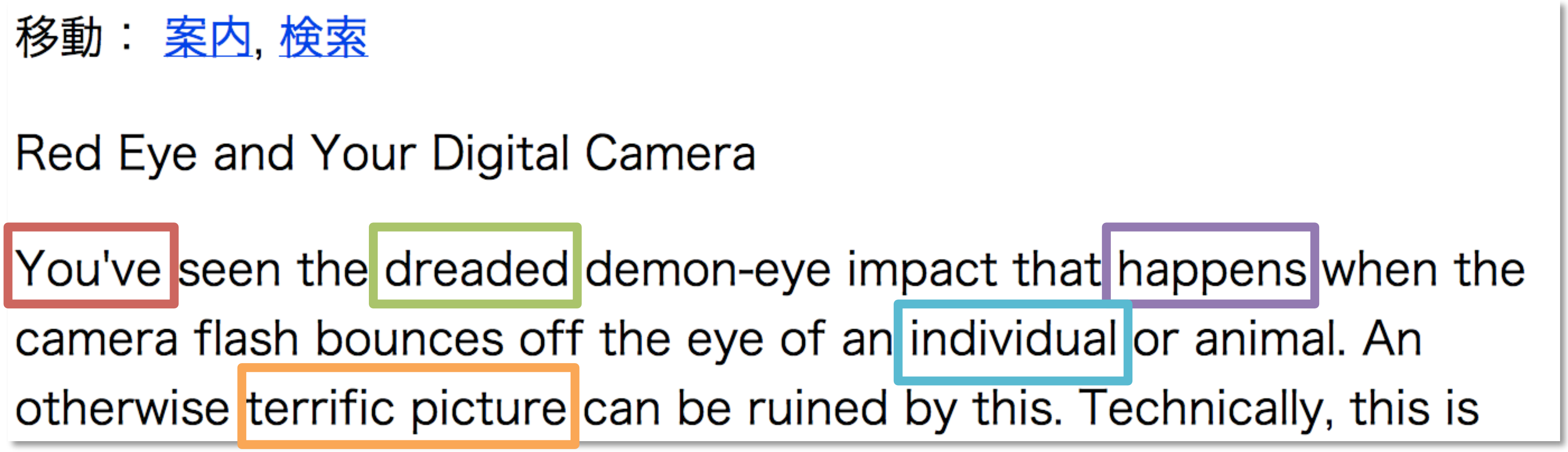
Spinning Example



Wechseln zu: Navigation, Suche

Red Eye and Your Digital Camera

You have actually seen the feared demon-eye impact that occurs when the camera flash bounces off the eye of a person or animal. An otherwise fantastic image can be ruined by this. Technically, this is



移動： 案内, 検索

Red Eye and Your Digital Camera

You've seen the dreaded demon-eye impact that happens when the camera flash bounces off the eye of an individual or animal. An otherwise terrific picture can be ruined by this. Technically, this is

Spinning Approaches

Human Spinning

- Hire a real person from an online marketplace (i.e. Fiverr, Freelancer) to spin manually
- Pros:
  - Reasonable text readability
- Cons:
  - Expensive ($2-8 / hr)
  - Not scalable (humans)

Automated Spinning

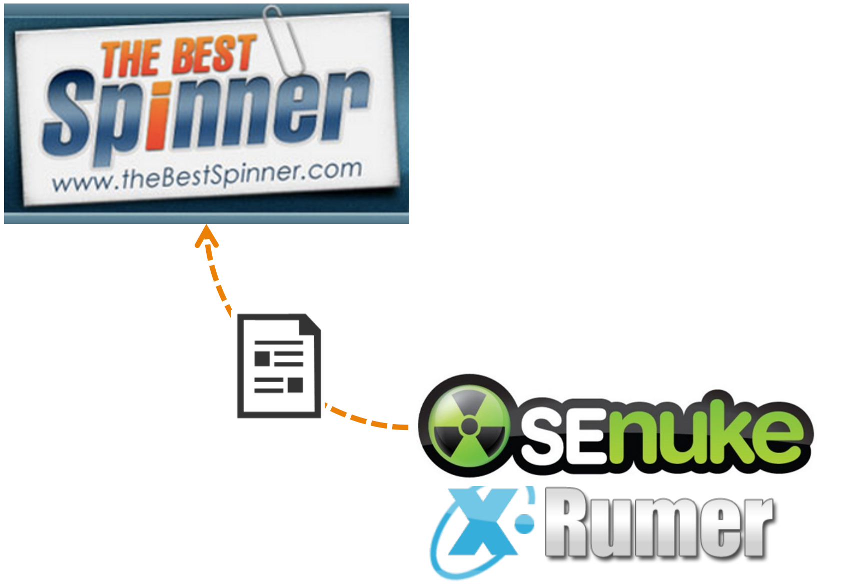
- Run software to spin automatically
- Pros:
  - Fast
  - Cheap ($5)
  - Scalable (500 articles / job)
  - Minimal human interaction
- Cons:
  - Can read awkwardly

Spinning in BHSEO

SEO Software

Start with a seed article
and SEO Software

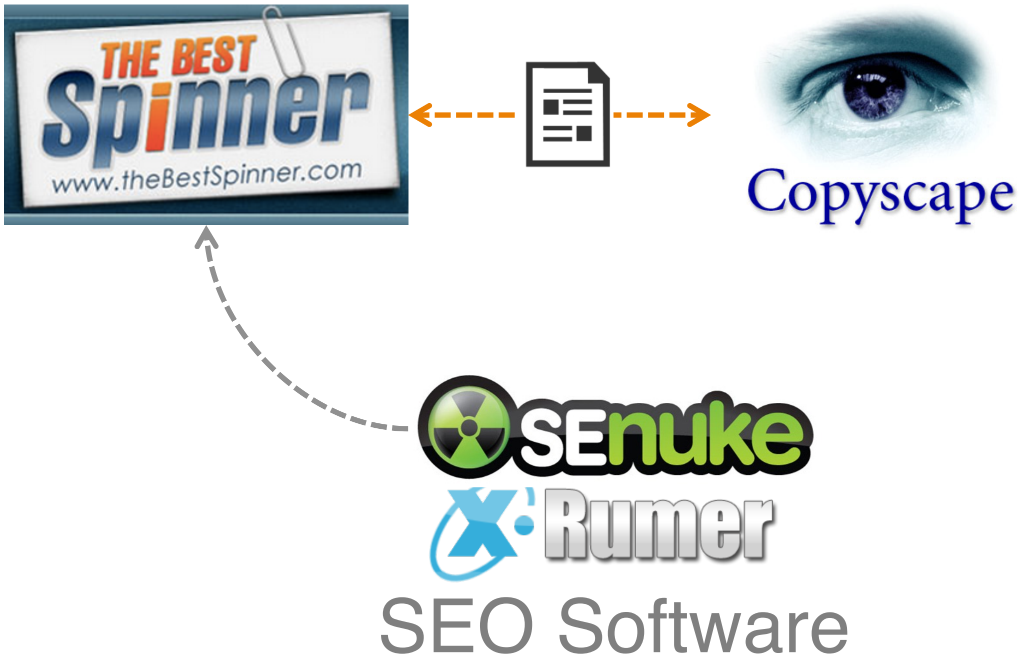
Spinning in BHSEO



SEO Software

SEO Software submits the article to spinner (TBS)

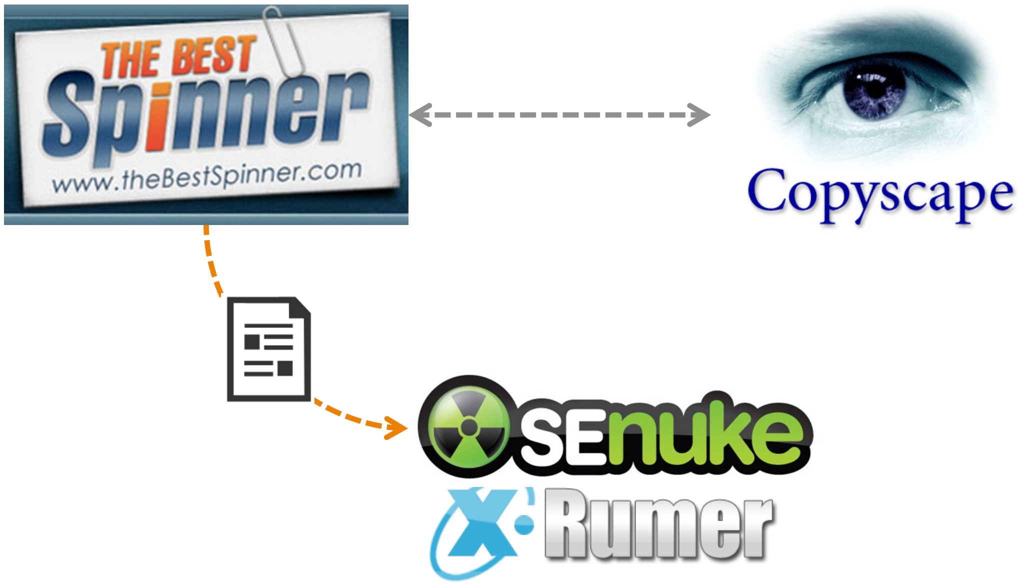
Spinning in BHSEO



SEO Software

TBS spins the article and verifies plagiarism detection fails

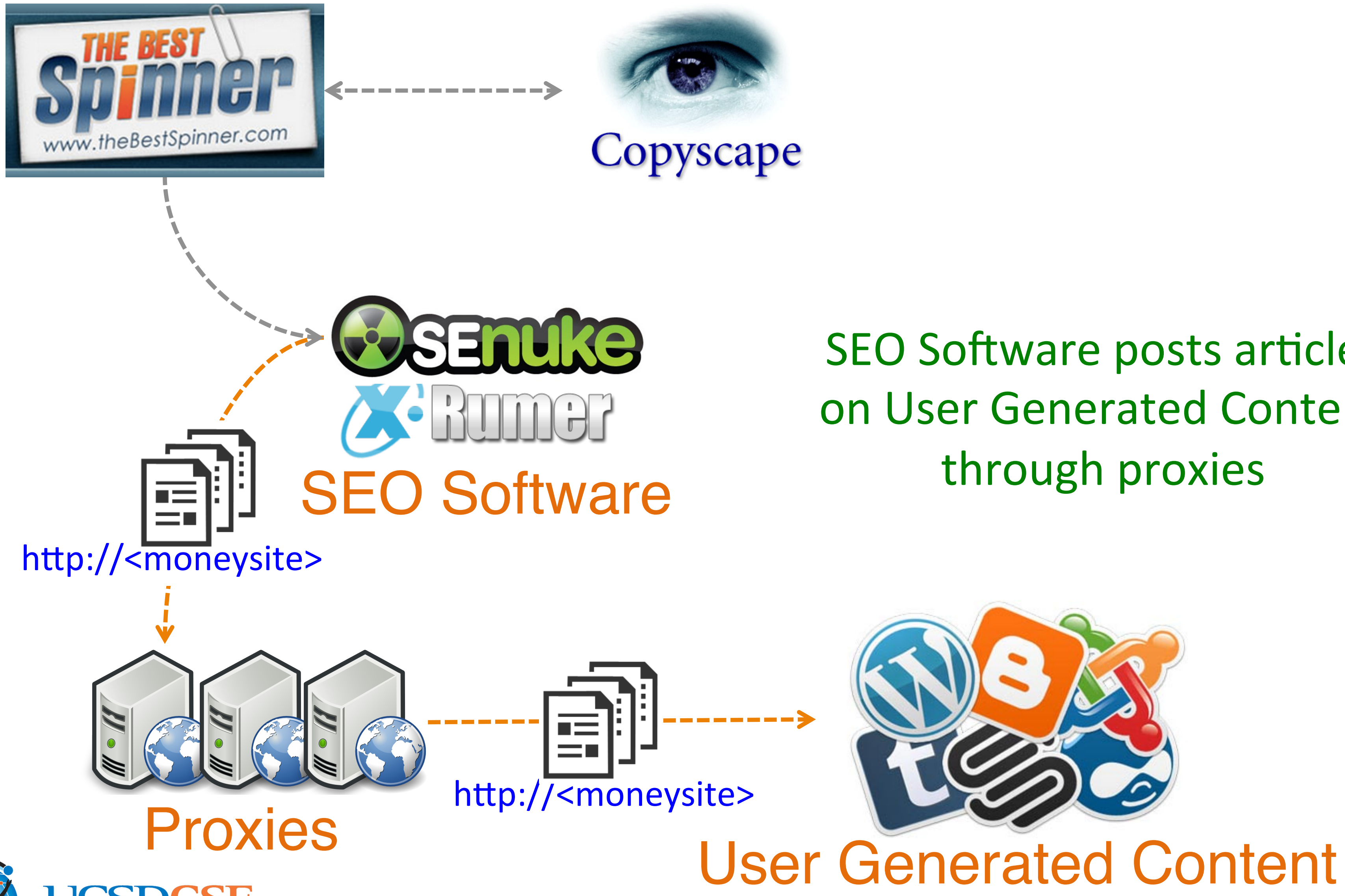
Spinning in BHSEO



SEO Software

SEO Software receives spun article

Spinning in BHSEO



SEO Software

http://<moneysite>

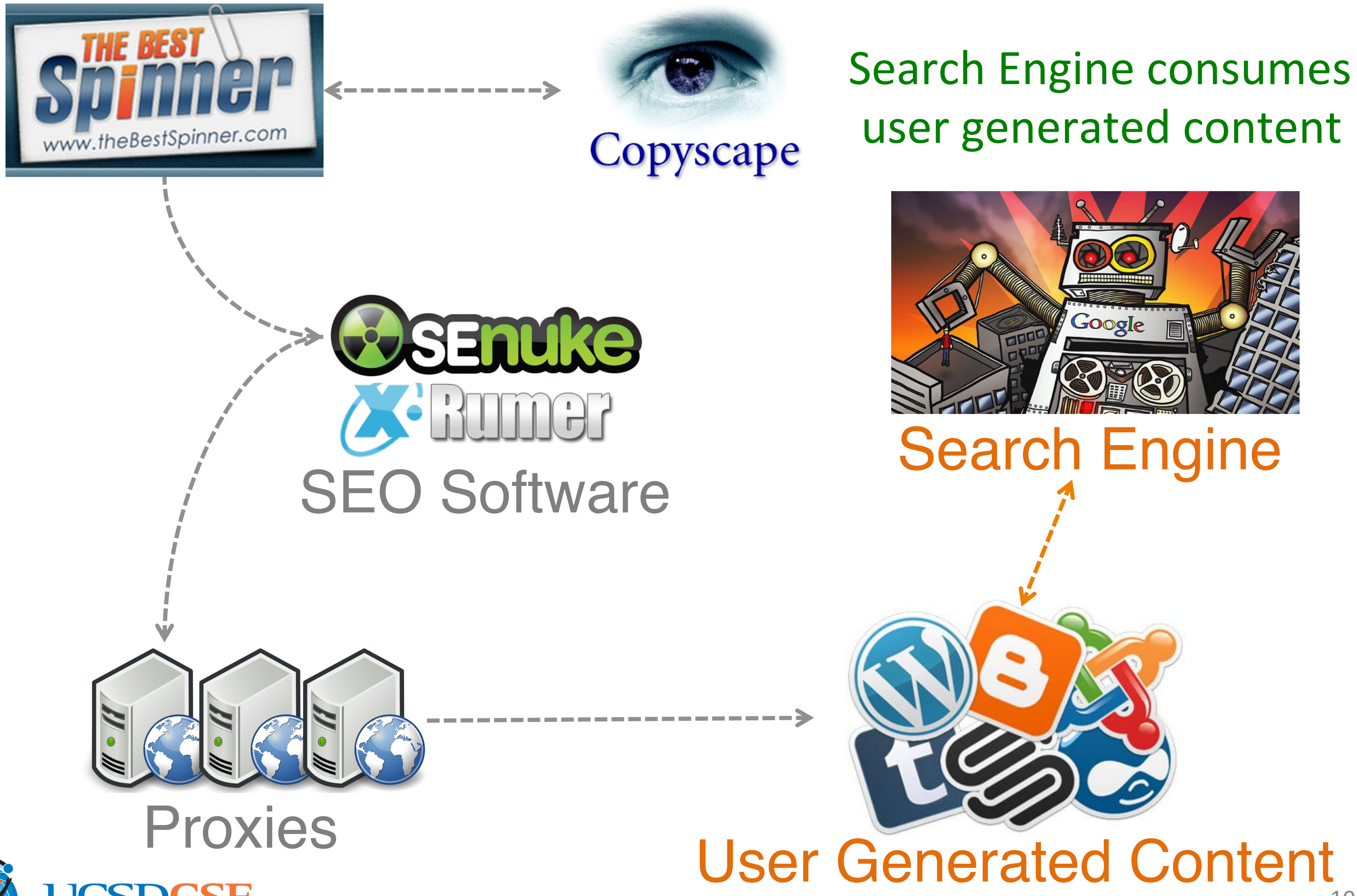
Proxies

http://<moneysite>

User Generated Content

SEO Software posts articles on User Generated Content through proxies

Spinning in BHSEO



Search Engine consumes user generated content

SEO Software

Search Engine

Proxies

User Generated Content

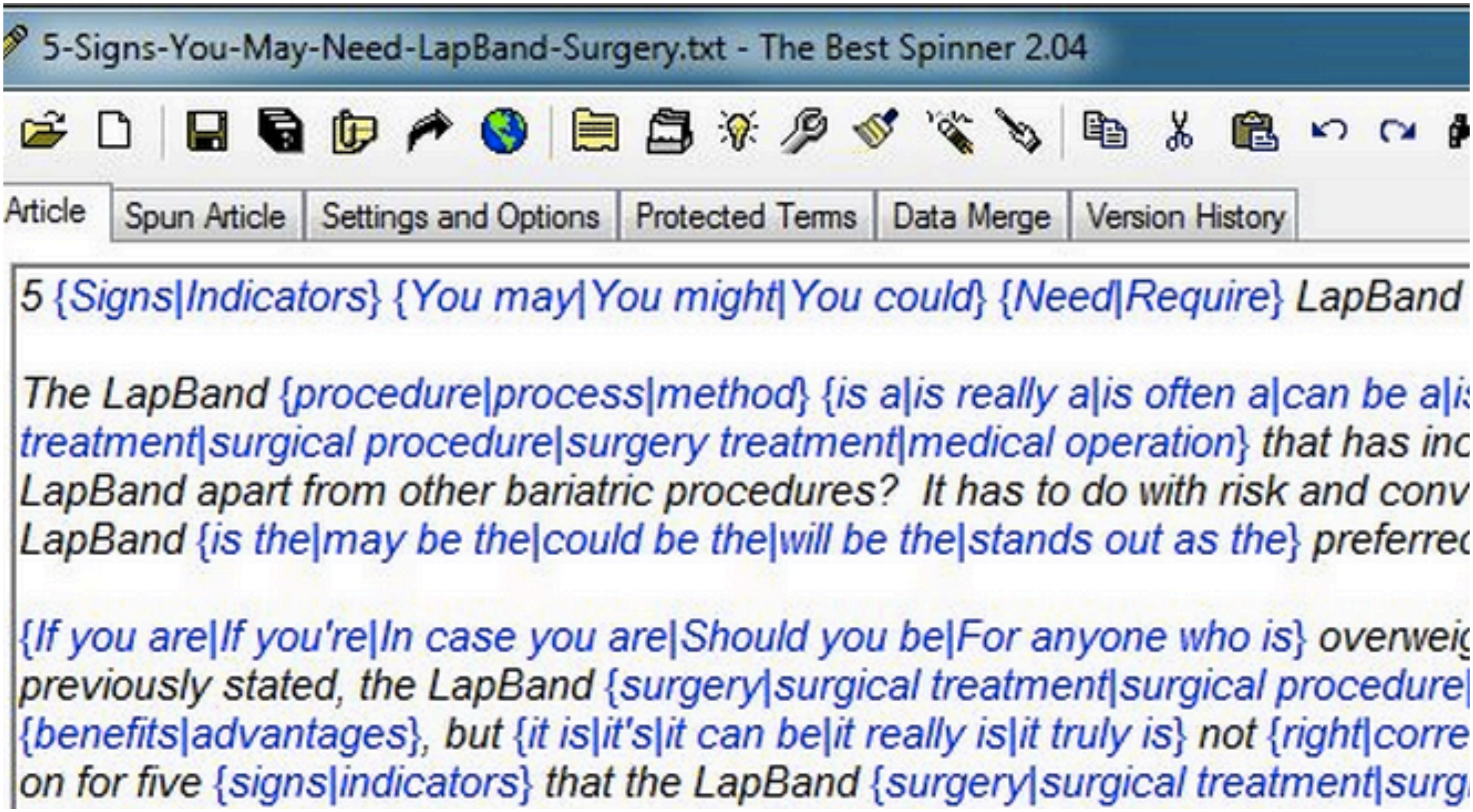
Goals

- Understand the current state of automated spinning software using one of the most popular spinners (The Best Spinner)
- Develop techniques to detect spinning using immutables + mutables
- Examine spinning on the Web using Dspin, our system to identify automatically spun content

The **B**est **S**pinner (**TBS**)

- TBS consists of two parts
  - Program (binary): provides the *user interface*
  - Synonym dictionary: a *homemade, curated list of synonyms* that are updated weekly
- Replaces text with synonyms from dictionary
- We extract the synonym dictionary through reverse engineering the binary

TBS Example



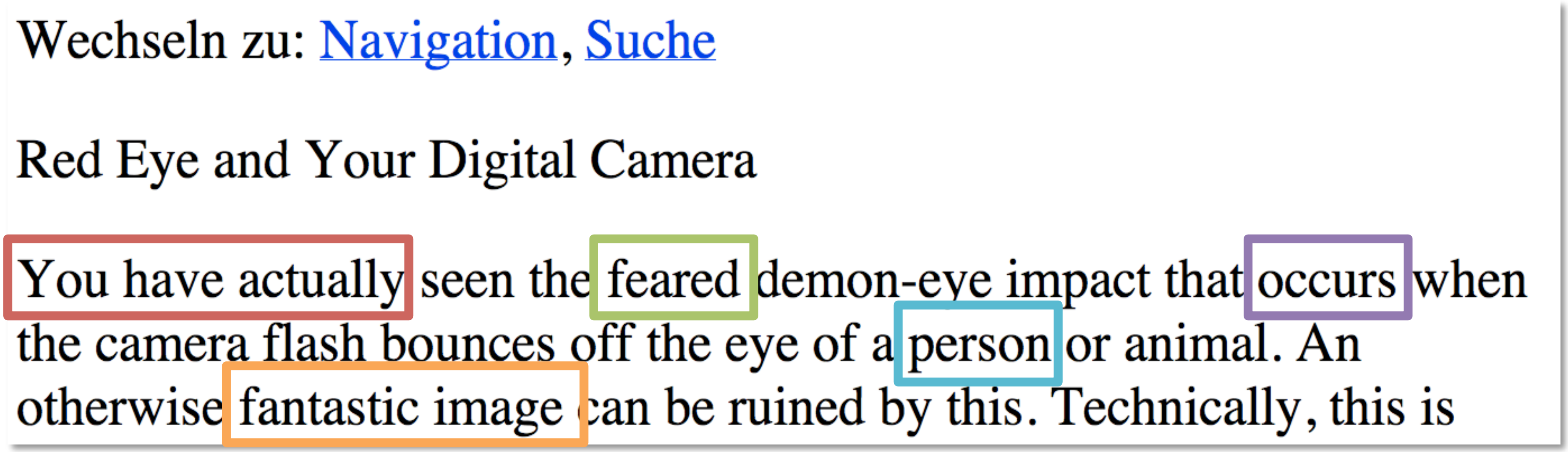
5 {Signs|Indicators} {You may|You might|You could} {Need|Require} LapBand

The LapBand {procedure|process|method} {is a|is really a|is often a|can be a|i
treatment|surgical procedure|surgery treatment|medical operation} that has inc
LapBand apart from other bariatric procedures? It has to do with risk and conv
LapBand {is the|may be the|could be the|will be the|stands out as the} preferre

{If you are|If you're|In case you are|Should you be|For anyone who is} overwei
previously stated, the LapBand {surgery|surgical treatment|surgical procedure|
{benefits|advantages}, but {it is|it's|it can be|it really is|it truly is} not {right|corre
on for five {signs|indicators} that the LapBand {surgery|surgical treatment|surg

Immutables + Mutables

- An article is composed of immutables *(NOT IN dictionary)* and mutables *(IN dictionary)*



Wechseln zu: Navigation, Suche

Red Eye and Your Digital Camera

You have actually seen the feared demon-eye impact that occurs when the camera flash bounces off the eye of a person or animal. An otherwise fantastic image can be ruined by this. Technically, this is

Spinning Detection Algorithm

- Immutables detection computes the ratio of shared immutables between two pages
  - Works well in practice except in corner case where there are few immutables to compare
- Mutables detection computes the ratio of all shared words after two levels of recursively expanding synonyms
  - Also works well and handles corner case, but expensive

Other Approaches

- Duplicate content detection is a well known problem for Search Engines
- Explored other approaches:
  - Hashes of substrings [Shingling]
  - Parts of speech [Natural Language Processing]
- Spinning is designed to circumvent these approaches (i.e. replace every Nth word, synonym phrases)

Validation

- Setup controlled experiment using TBS
- 600 article test data set
  - Started with 30 seed articles
    - 5 articles from 5 different article directories
    - 5 articles randomly chosen from Google News
  - Each article spun 20 times w/ bulk spin option
- Immutables detects all spun content and matches with the source

DSpin

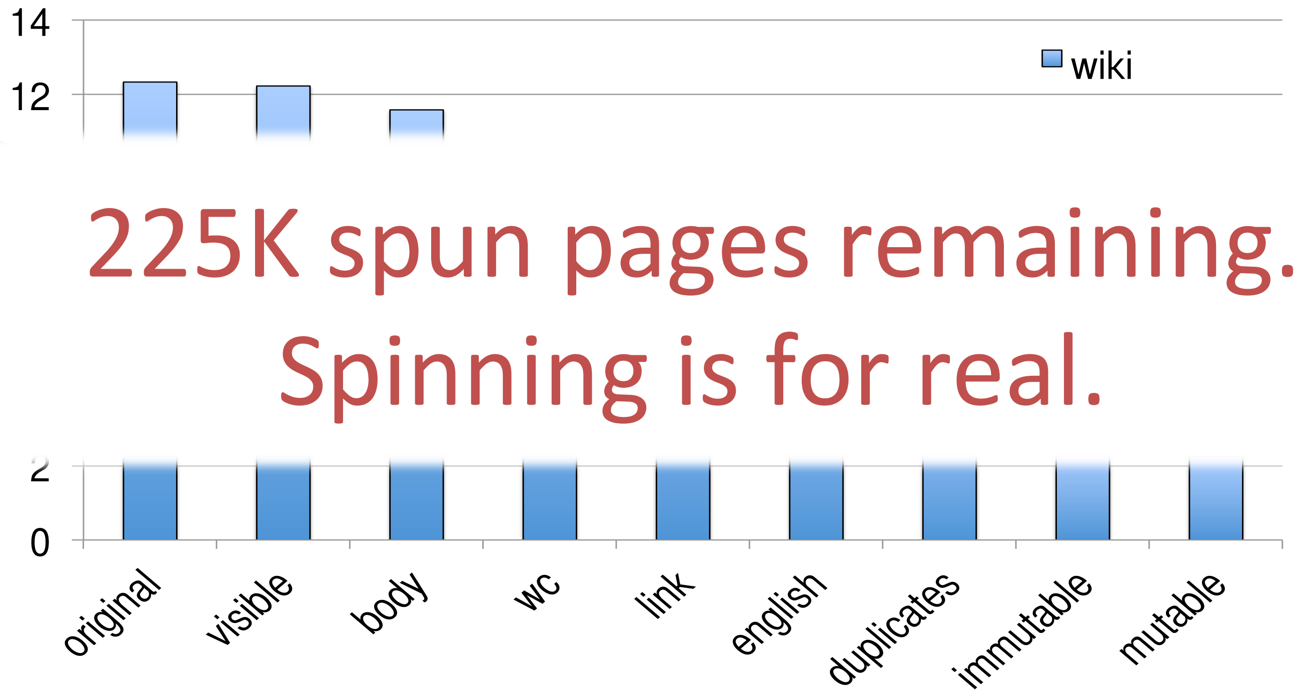
- Detection from Search Engine POV
  - Input: set of article pages crawled from the Web
  - Output: set of pages flagged as auto spun
- Build graph of clusters of “similar” pages using immutables + mutables approach
  - Each page represents a node
  - Create edges between pairs of nodes using immutables, verify edges using mutables
  - Each connected components is cluster

Results

- Ran DSpin on a real life data set
  - Set of 797 abused wikis
  - Crawl each wiki daily for newly posted articles
  - Collected 1.23M Articles from Dec 2012
- Address the following questions:
  - Is spinning a problem in the wild?
  - Can we characterize spinning behavior?

Filtering

- Filter out pages that are: non-English, exact duplicates, < 50 words, or primarily links



225K spun pages remaining.

Spinning is for real.

Wiki Content

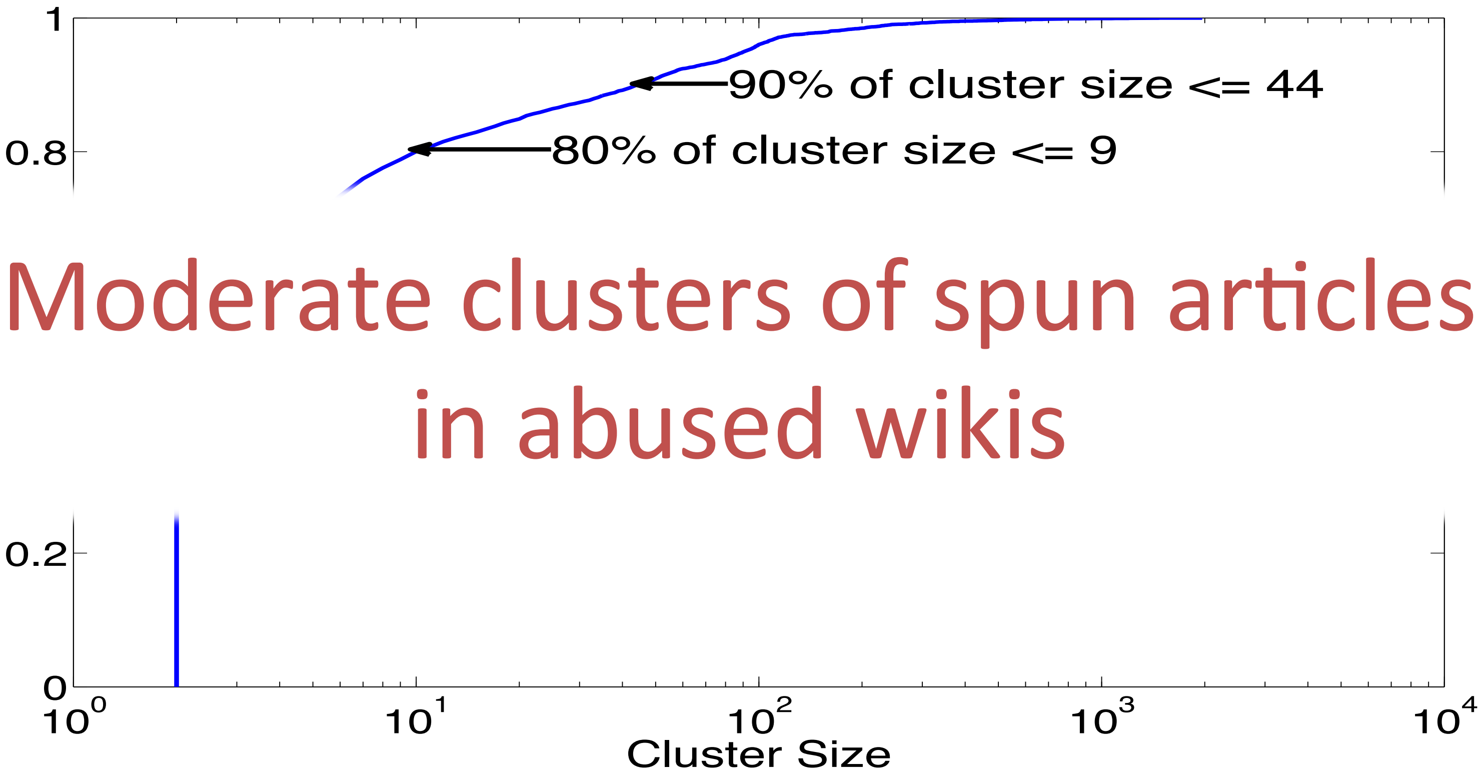
skin
health
credit
go
needs
bsite
assist
world
energy
plan
give
computer
customers
art
discover
tion
life
course
nerous
contact
form
care
medical
price
create
diet
water
try
day
npany
name
pay
result
come
software
see
another
start
methods
means
consider
looking
ystem
newsletter
video
check
service
free
offer
ing
made

Spinning campaigns target business + marketing terms

design
family
area
services
working
organization
cost
set
increase
got
position
program
insurance
online
items
choose
feel
sort
deal
understand
visit
read
include
think
save
market
hair
support
require
allow
page
hostin
treatment
couple
used
provides
yourself
experience
during
cleaning
fact
customer
ensure
folks
loss
order
home
personal
content
person
years
reason
mind
car
purchase

Cluster Size

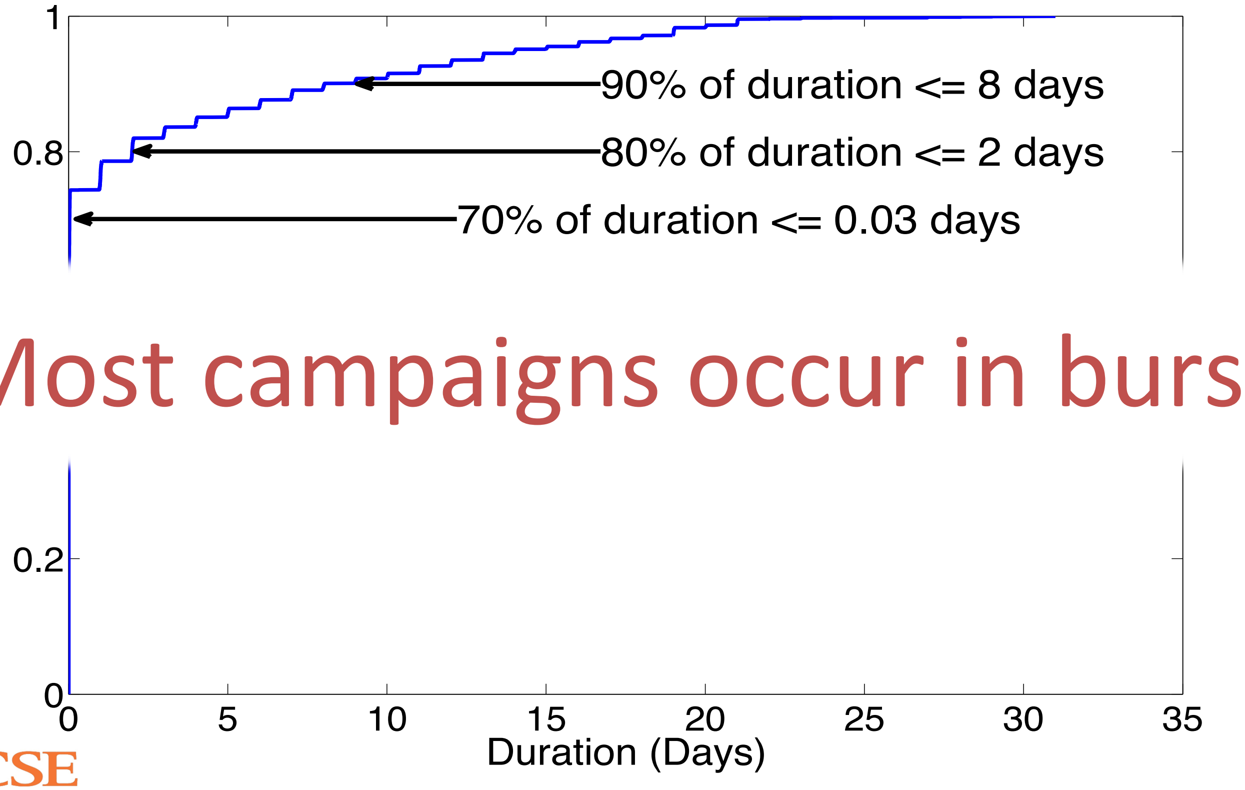
- 12.7K clusters from 225K spun pages



Moderate clusters of spun articles in abused wikis

Timing Duration

- Duration reveals how long a campaign lasts
- Compute by extracting dates, max – min



Most campaigns occur in bursts.

Conclusion

- Proposed + evaluated a spinning detection algorithm based on immutables + mutables that Search Engines can implement
- Demonstrated the algorithm's applicability on a real life data set (abused wikis)
- Characterized the behavior of at least one slice of the Web where spun articles thrive

Thank You!

- Q&A

TBS Coverage

- Only one synonym dictionary was used to implement DSpin, is this system still applicable widely (i.e. for other spinners)?
  - We had no prior knowledge about how articles from abused wikis were spun
  - Yet we still detected spun articles

Synonym Dictionary Churn

- How much does the synonym dictionary change over time?
  - We re-fetched synonym dictionary four months after the initial study and found that 94% of terms remain the same
  - Furthermore, DSpin detected spun articles posted months prior

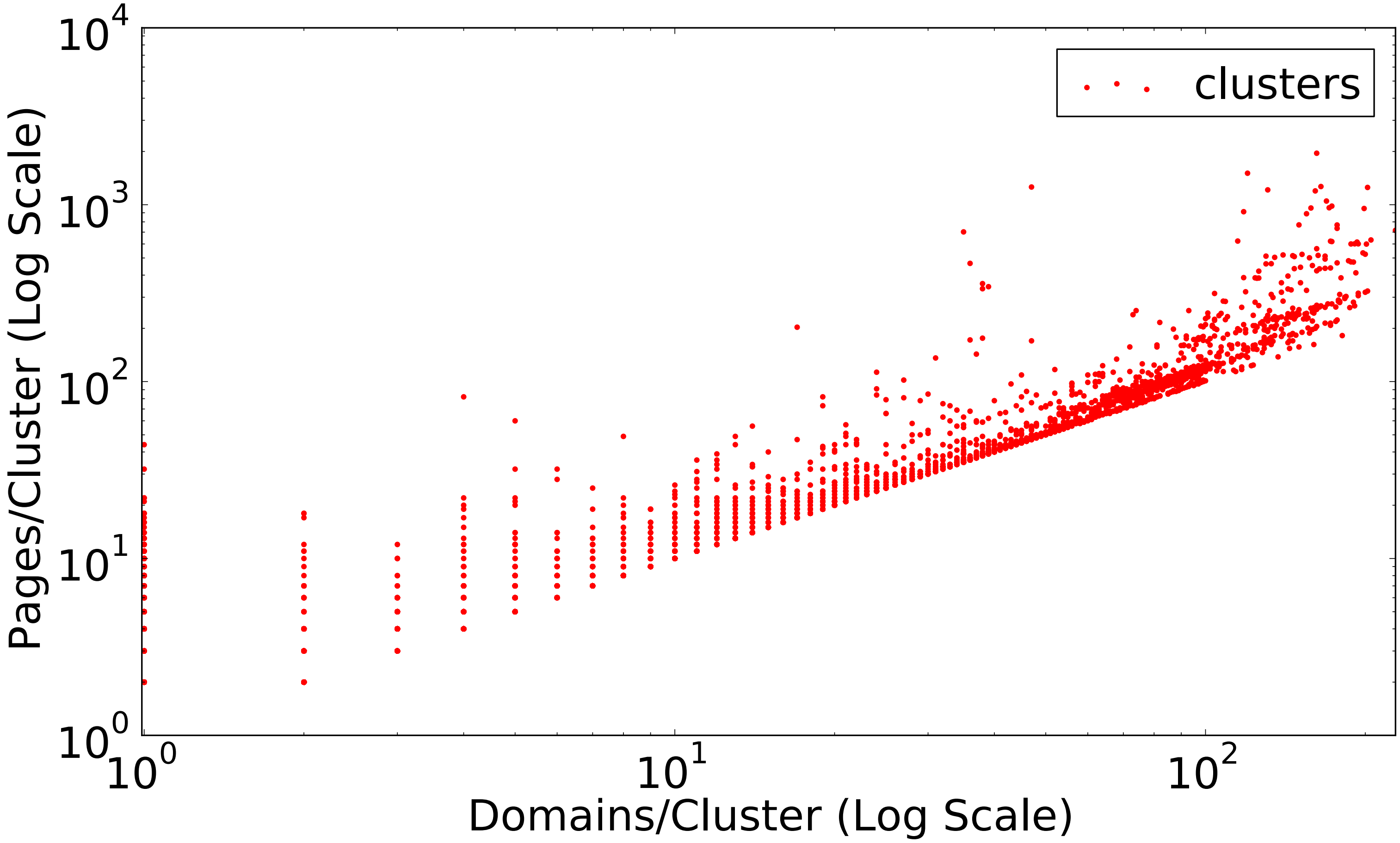
Synonyms in the Cloud

- What if the spinner stores the synonym dictionary in the cloud?
  - There is an operational cost for the spinner (network bandwidth == $$$)
  - Can still reconstruct synonym dictionary through controlled experiments (i.e. submitting our own articles for spinning)

Scalability

- How can Search Engines implement the immutables algorithm?
  - Assume Search Engines already perform duplicate content detection
  - Can think of immutables approach as performing duplicate content detection on the immutables portion of the pages (a subset of what is already currently done)

Pages/Cluster vs. Domains/Cluster

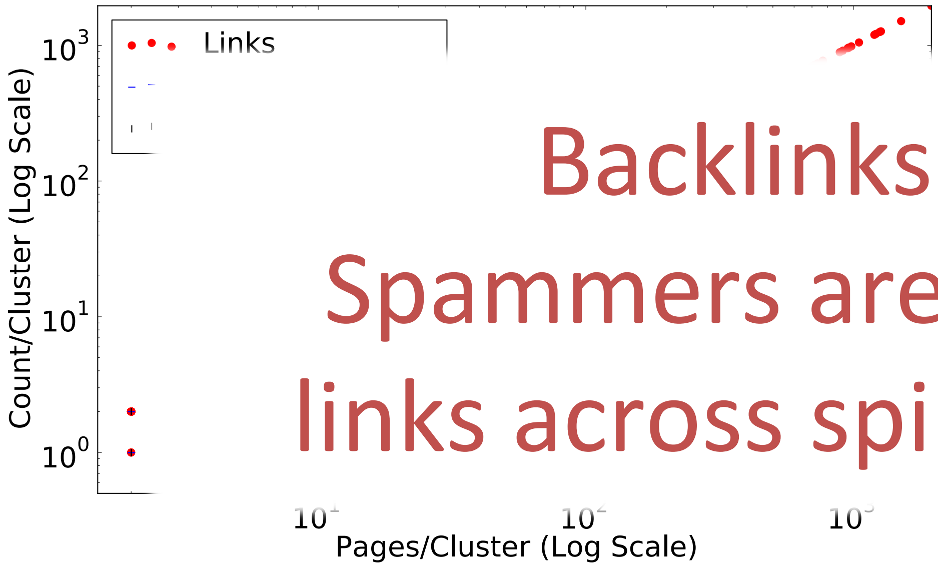


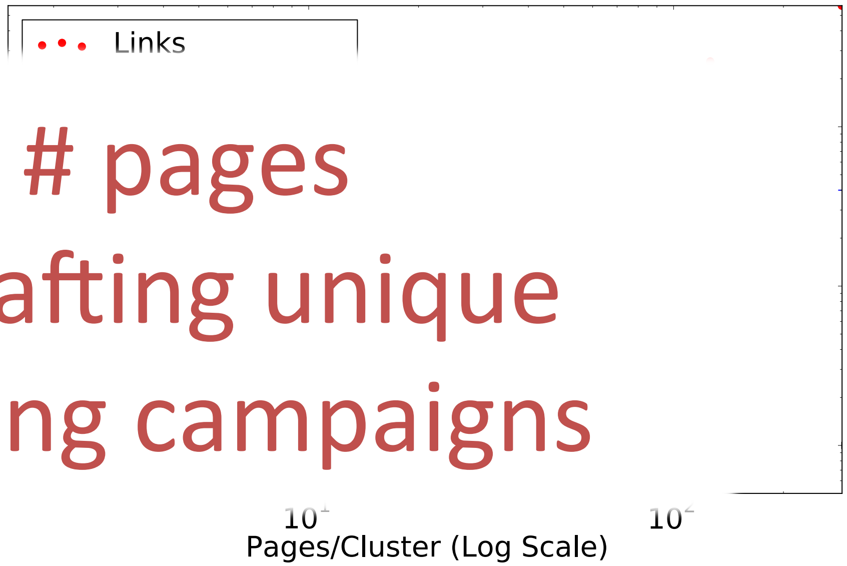
UCSDCSE
Computer Science and Engineering

Backlinks

Wiki

GoArticles

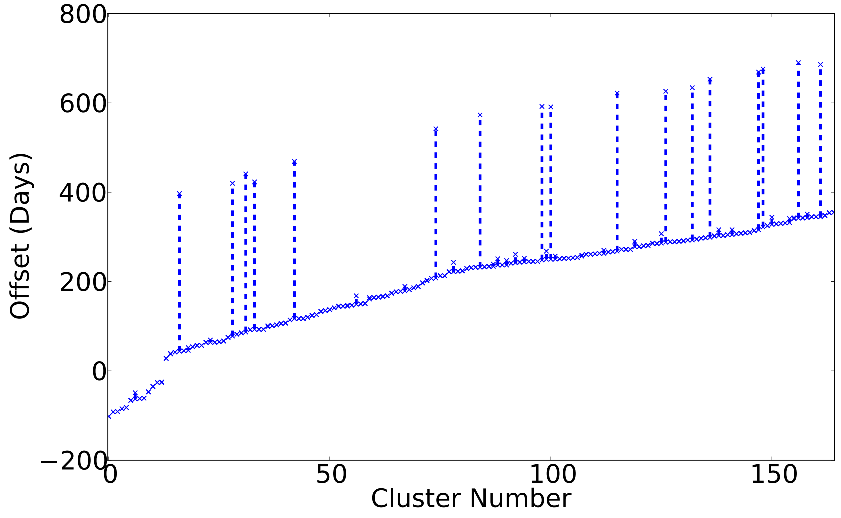


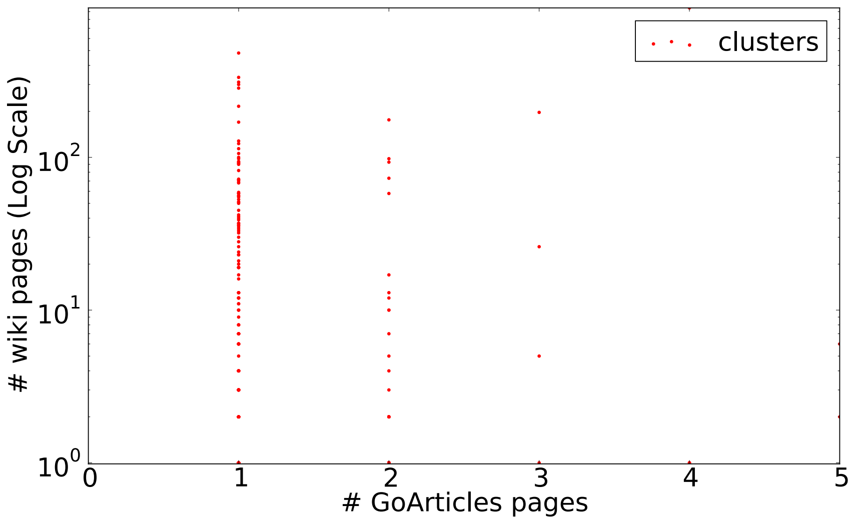


Backlinks ~= # pages

Spammers are crafting unique links across spinning campaigns

Seed Page

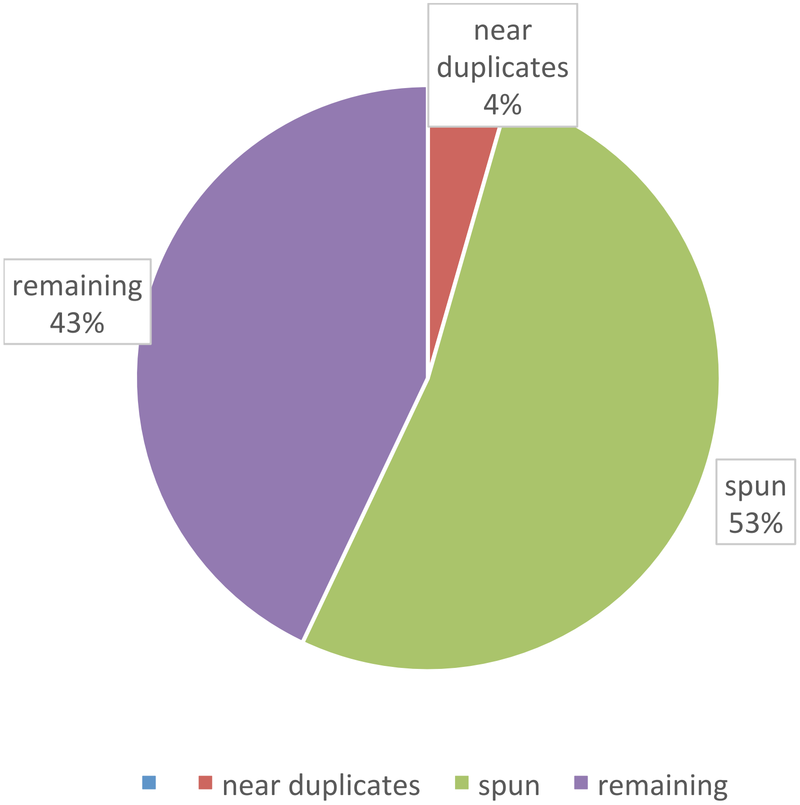




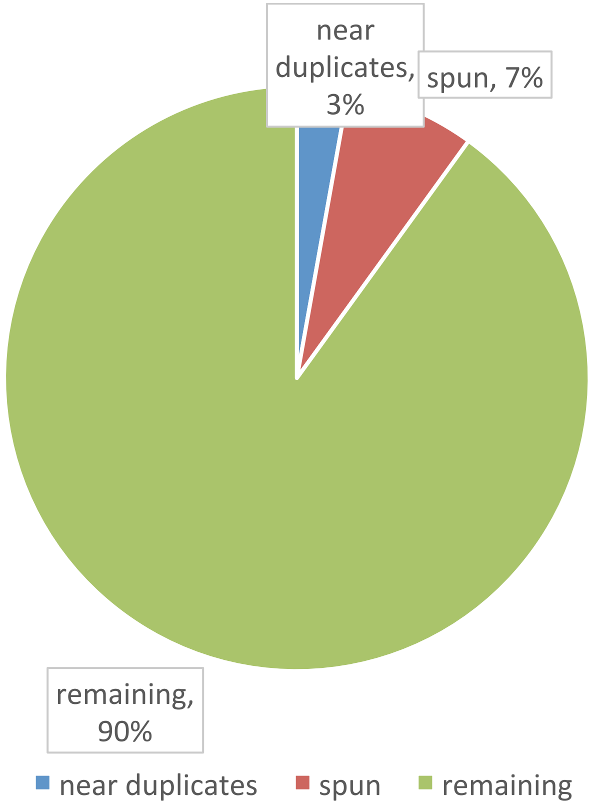
UCSDCSE
Computer Science and Engineering

Breakdown

Wiki Dataset

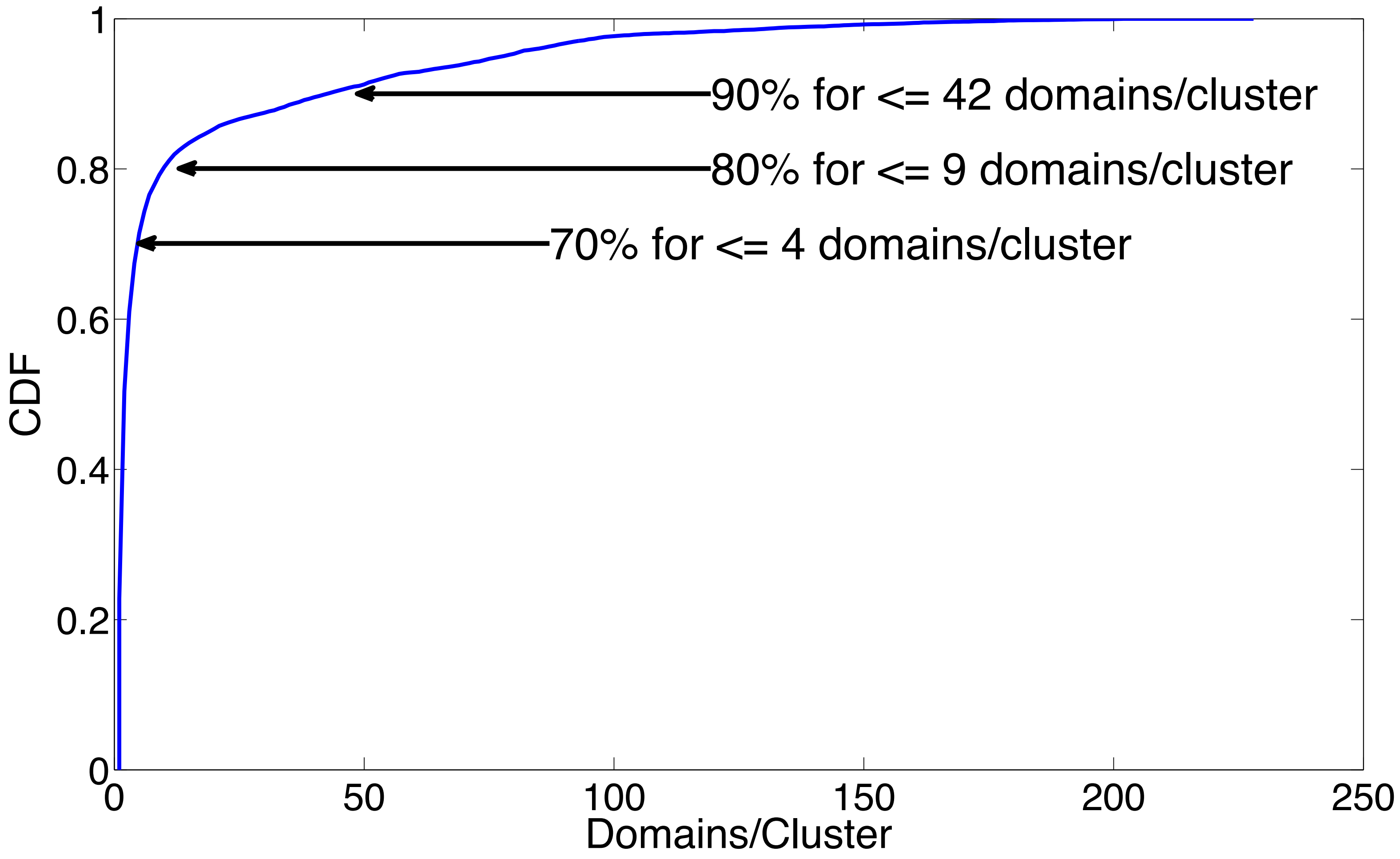


GoArticles



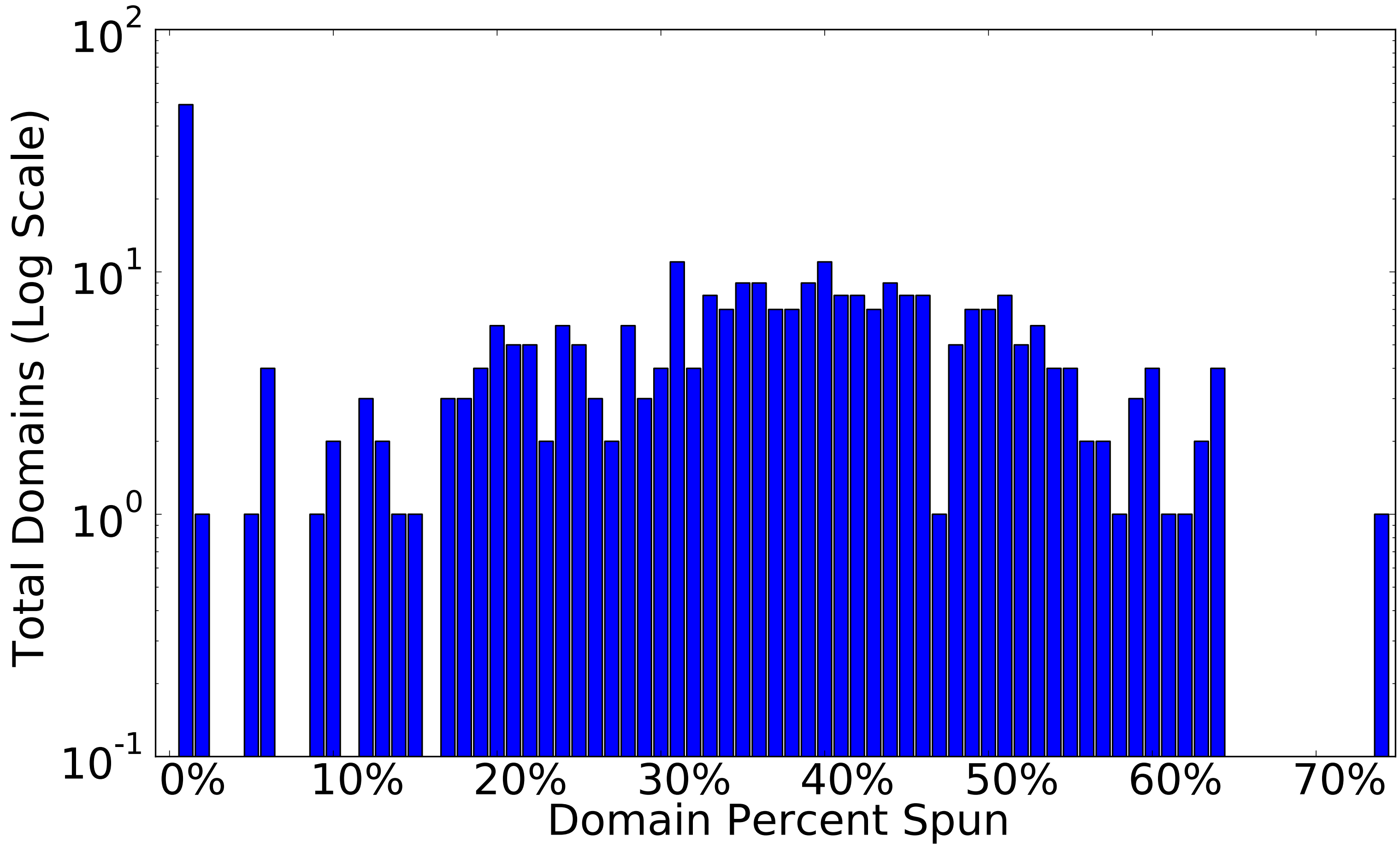
UCSDCSE
Computer Science and Engineering

Domains/Cluster



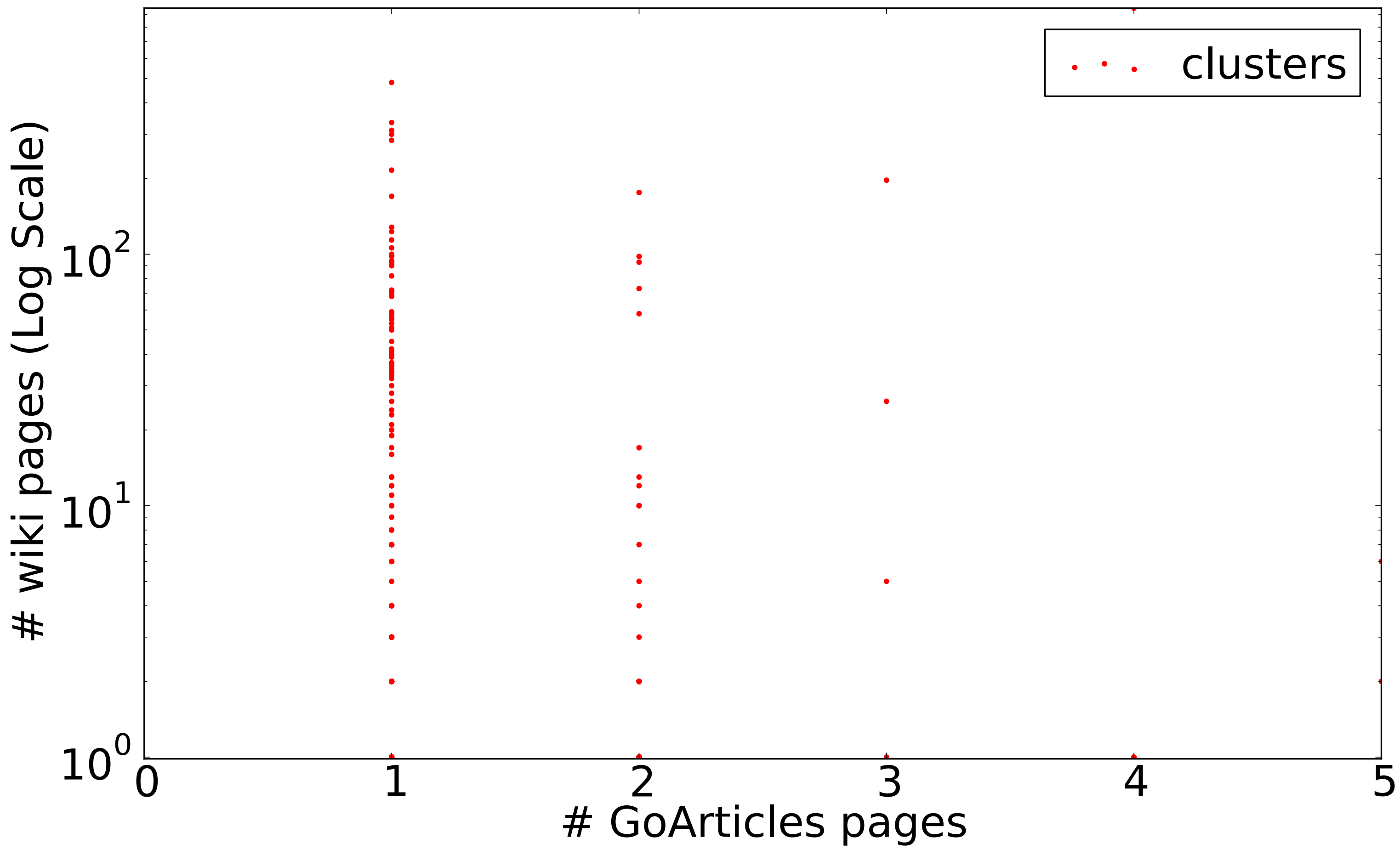
UCSDCSE Computer Science and Engineering

Percent Spun per Domain



UCSDCSE
Computer Science and Engineering

Seed Pages continued



UCSDCSE
Computer Science and Engineering