



CSCI 15: AN INTRODUCTION TO THE MODERN INTERNET

Lecture 9: Search Engines, the Structure of the Web, and the
Deep Web

ADMIN

- Papers are due Friday!!!
- As I mentioned in the first class: no extensions (my grades are due)

WIKIPEDIA GAME!

- How many clicks to get to Williams College?
- Start with Wikipedia page (given on website)
- Only click Wikipedia links
 - No dates, no “discussion” etc.
- What’s the shortest path you can find?
- Example: Harvard to Williams
- 20 minutes, 4 pages on the website (plus random generator)

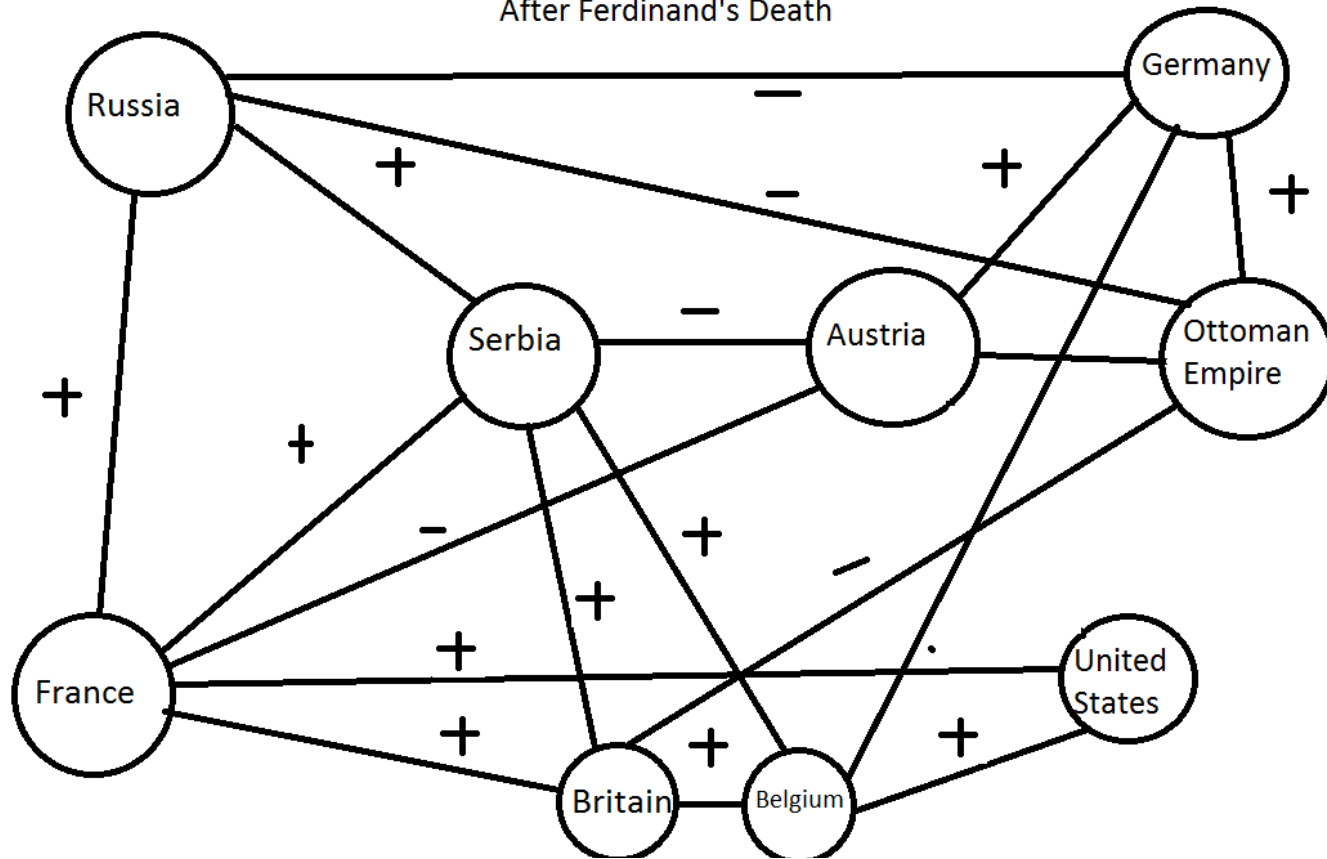
WIKIPEDIA GAME

- What are we drawing here?
- Does this idea apply to other topics?
- What kind of pages were most helpful?

5 DEGREES TO KEVIN BACON

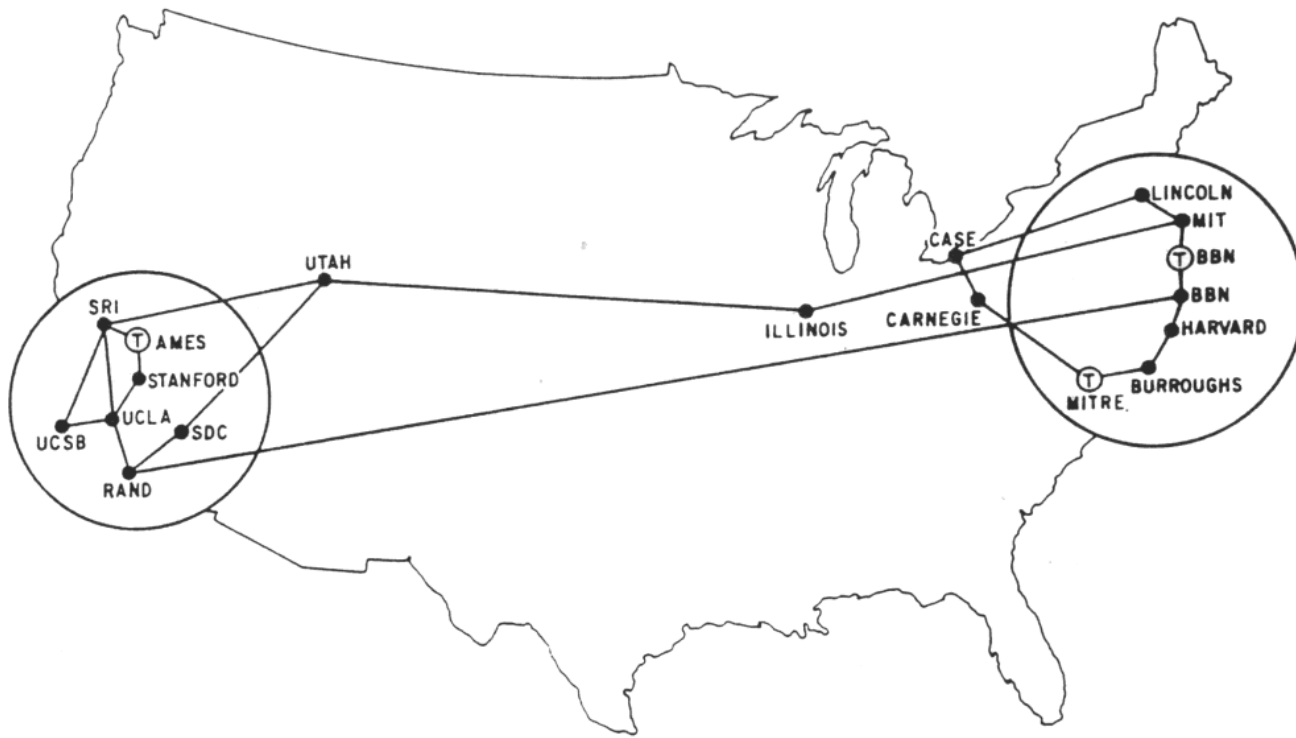
- <https://oracleofbacon.org/>
- Get from one actor to Kevin Bacon, where links are due to appearing in a common movie

After Ferdinand's Death



- Way to represent relationships
- Friendship graph, network/web graph

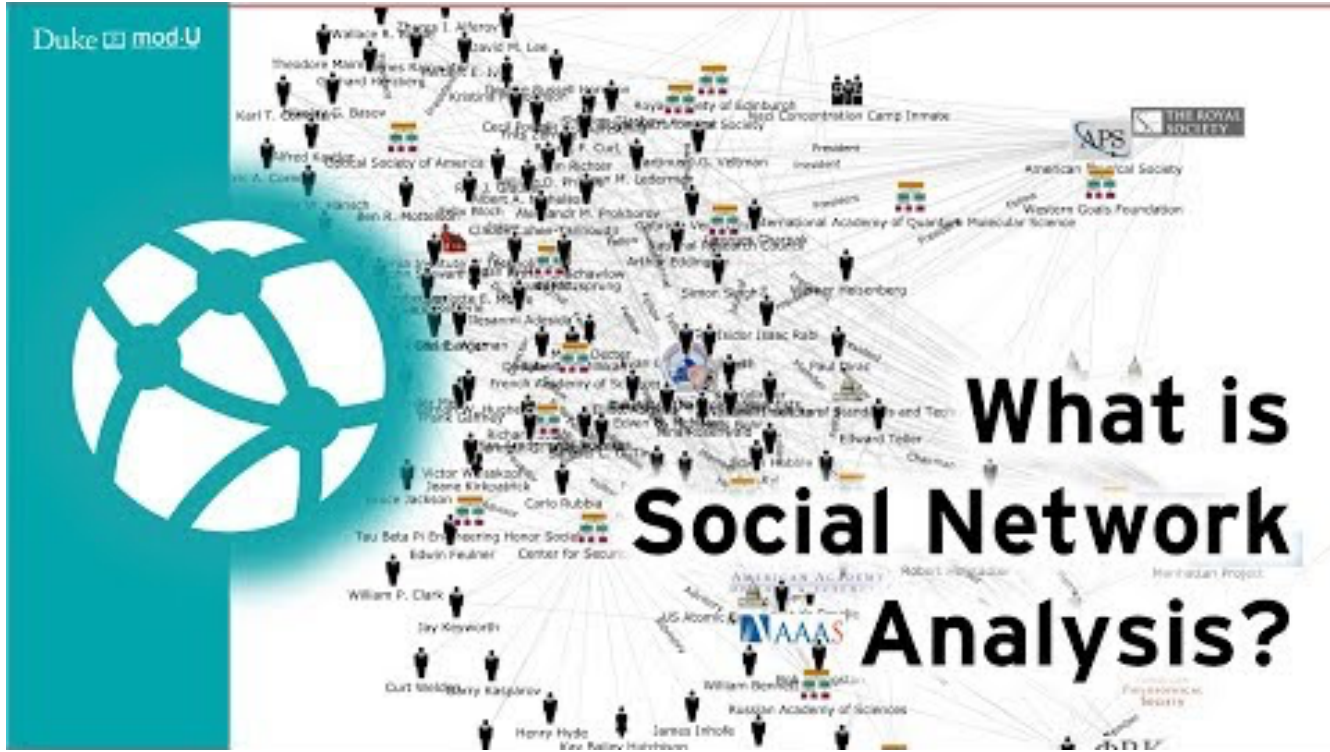
GRAPH



MAP 4 September 1971

GRAPH

- Way to represent relationships
- Friendship graph, network/web graph



- Way to represent relationships
- Friendship graph, network/web graph

GRAPH

BRINGING THESE TOGETHER

- What does the Wikipedia game (or Kevin Bacon game) mean in terms of a graph?
- “Path” of edges from one node to another



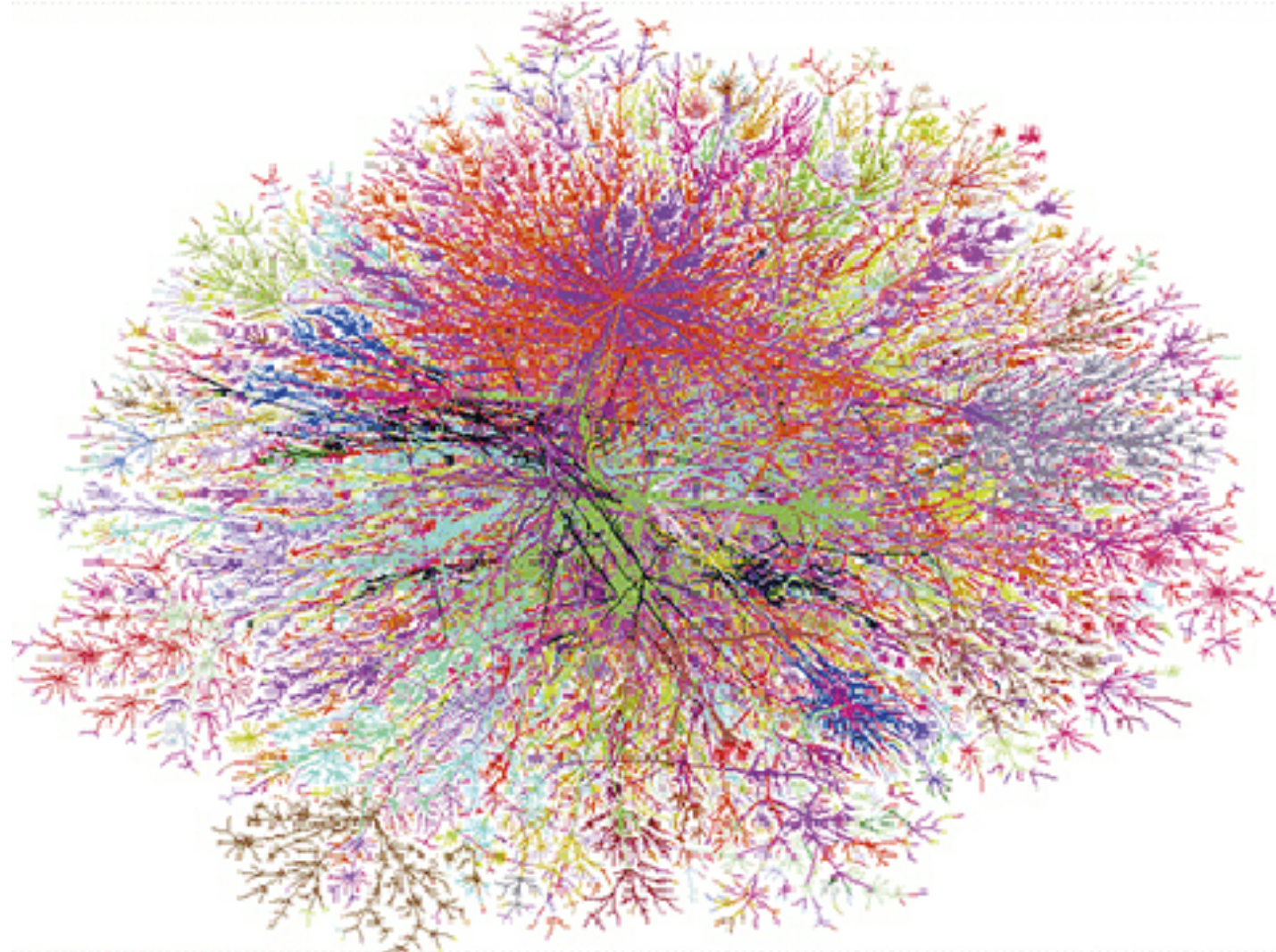
WEB GRAPH

live.com
What can we learn from this?

FUN WITH THE WEB GRAPH

- <http://internet-map.net/>
- Links not shown (oh well)
- Why are some websites large?

MAP OF THE WEB IN 1998



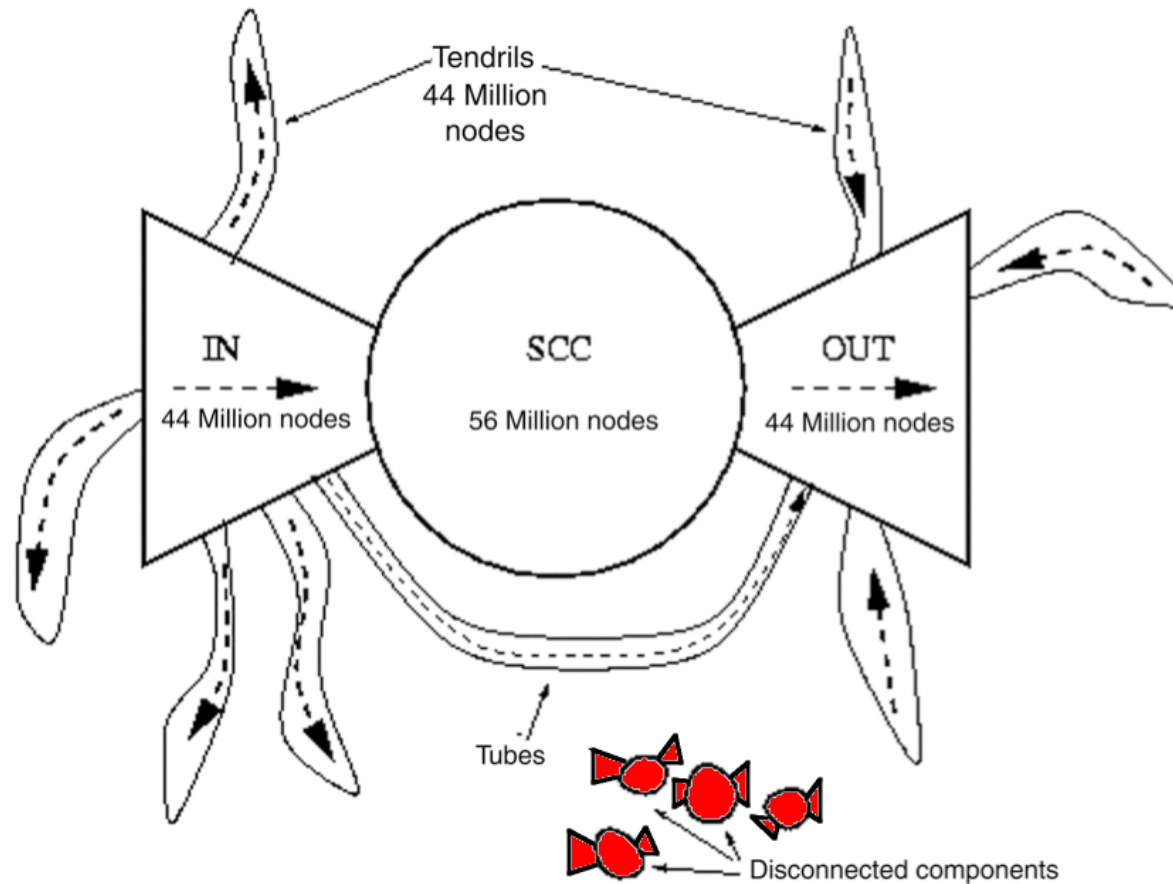
SHAPE OF THE WEB

- What does this look like?
- Random?
- Lots of big sites that get linked to, but don't link to anyone else?

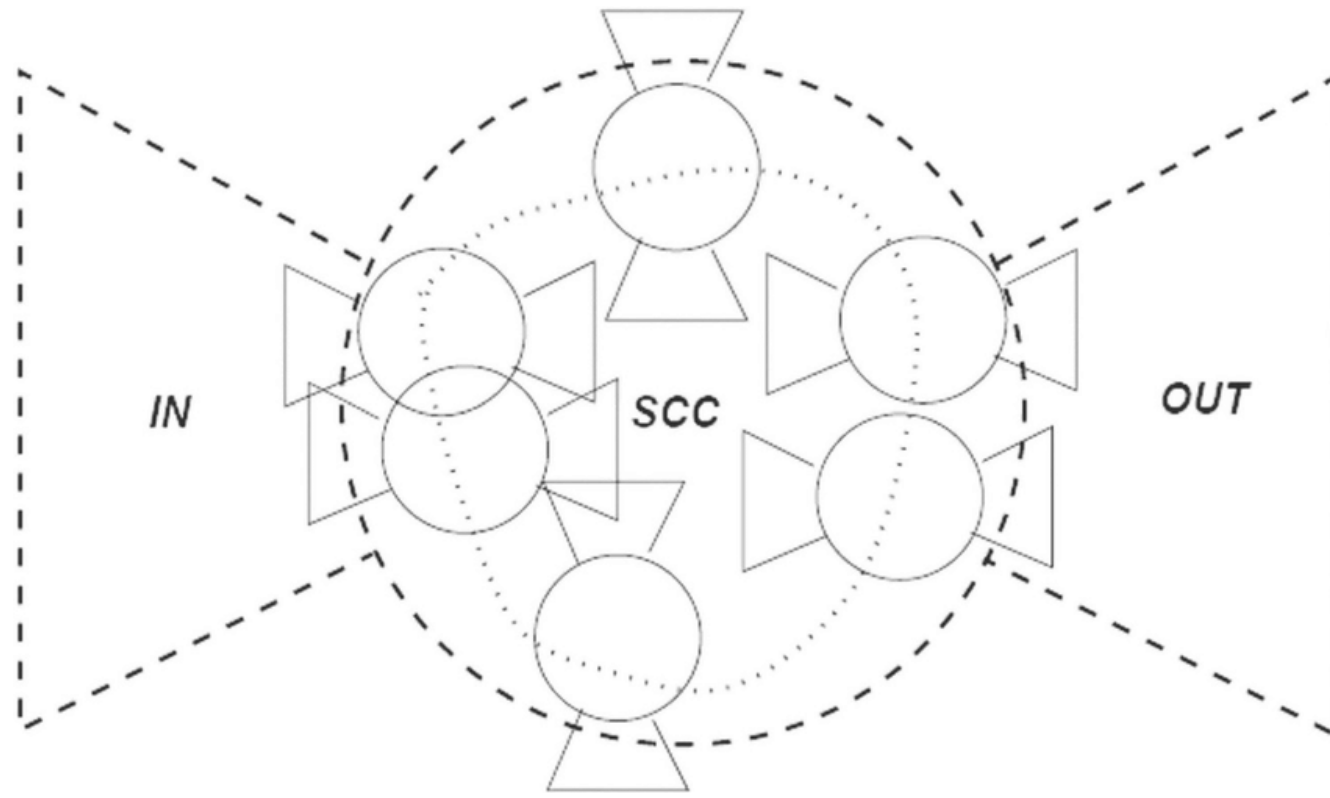
SHAPE OF THE WEB

- Bowtie!
- Middle portion: big webpages that all have a path to each other
 - The “Kevin Bacon” portion
- Left portion: links into the middle portion
- Right portion: links *from* the middle portion

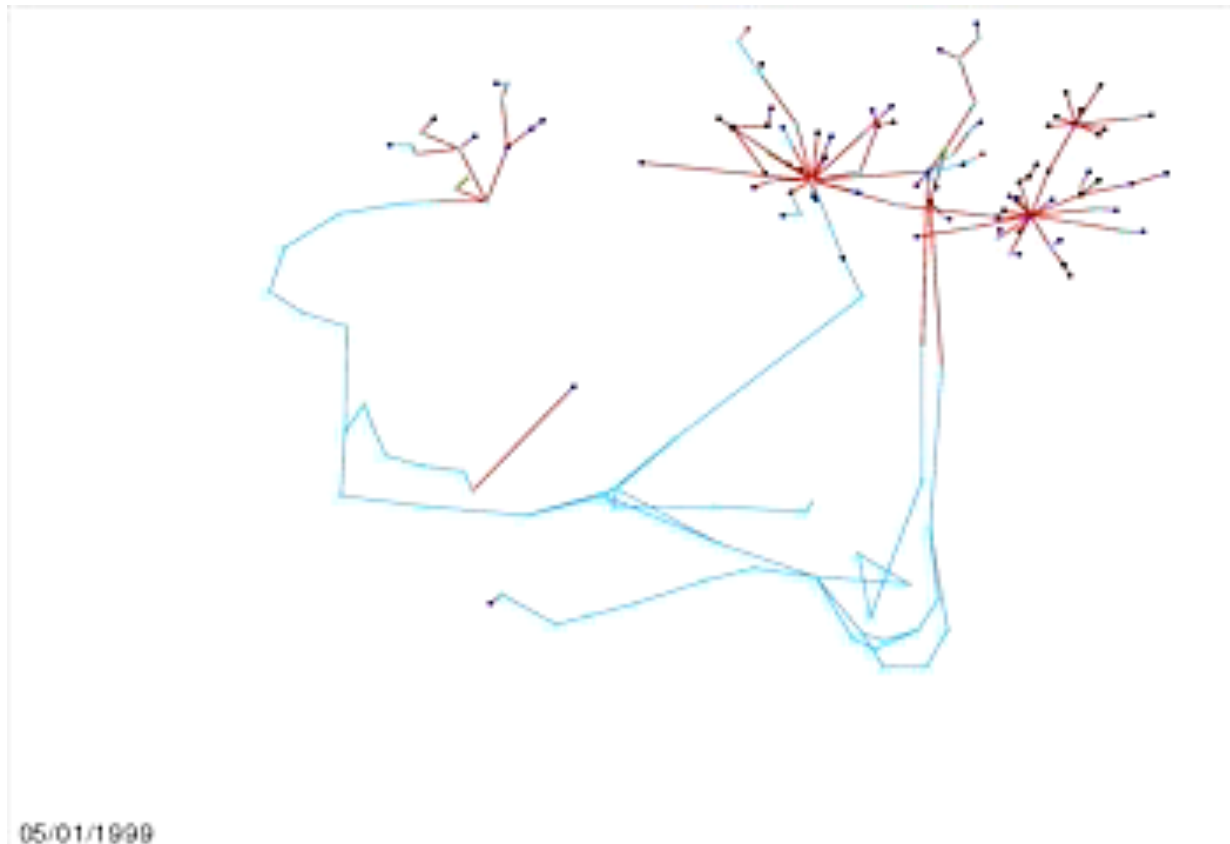
SHAPE OF THE WEB (BRODER ET AL. 2000)



MODERN OUTLOOK



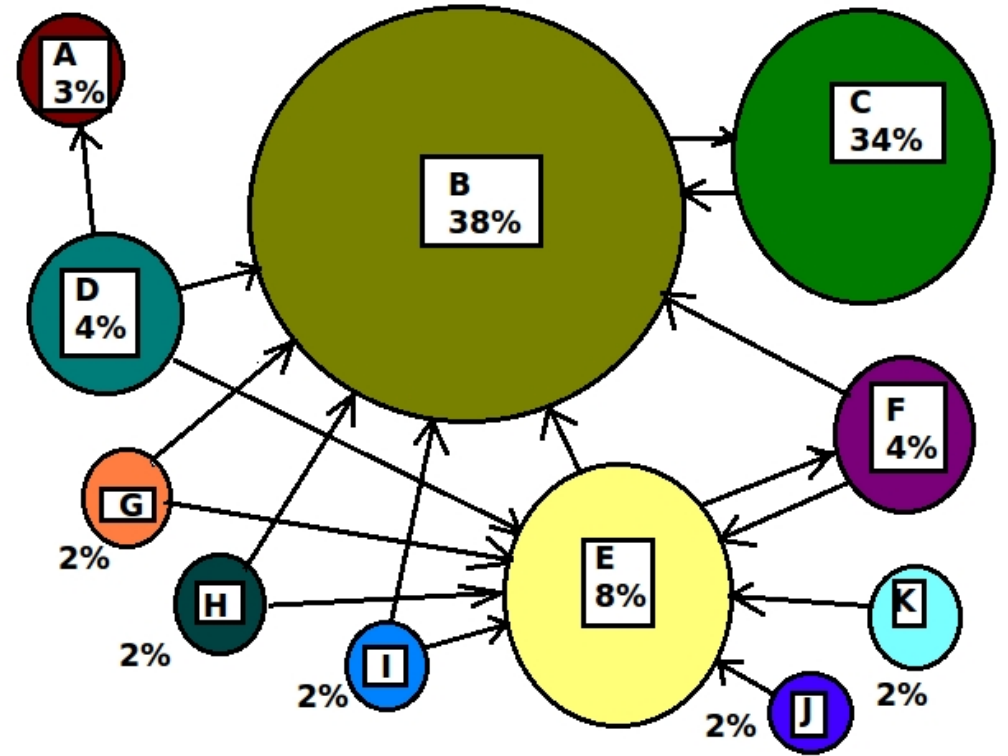
STRUCTURE CAN TELL A STORY



The internet in Yugoslavia in May 1999; each frame is one day. From “The effects of war on the Yugoslavian Network.” Steven Branigan and Bill Cheswick, 1999.

WEB SEARCH

- You're google and it's 1998
- What pages are "important"?
- What pages are not important?



PAGERANK INTRO



PAGERANK INTRO - DISCUSSION

- How do we rank importance of webpages?
- What does importance come from?
- What are some problems with this approach?

PAGERANK FOLLOWUP

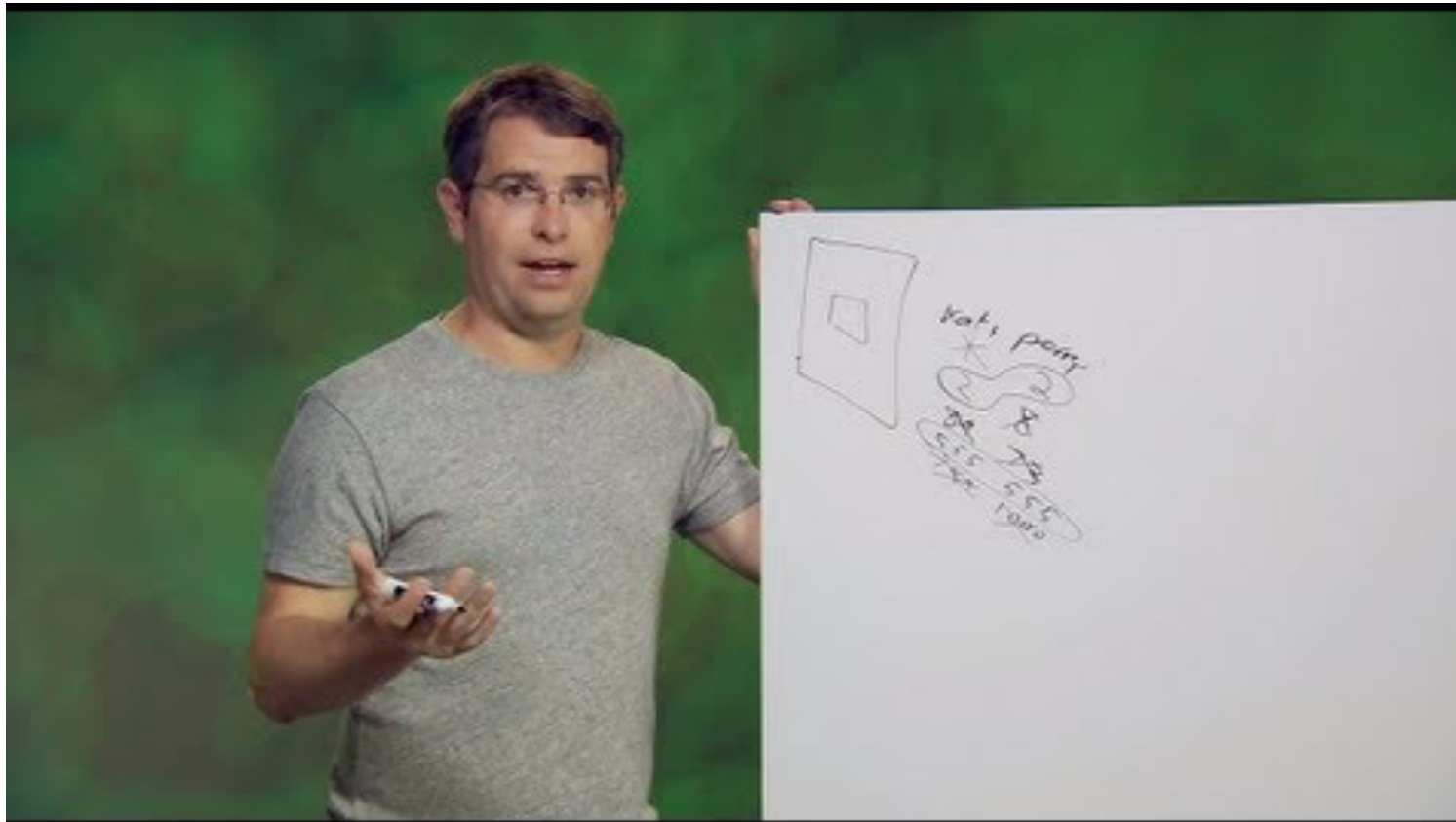
- Can calculate pagerank somewhat quickly using linear algebra
- Get “importance” based on web structure alone



DESIGNING A SEARCH ENGINE

- You're Larry Page and Sergey Brin, and you have this great algorithm for calculating importance
- How do you go from there?

HOW GOOGLE WORKS



WHAT'S WRONG WITH THIS?

- Why do search engine rankings matter?
- Can Google as we've discussed it be abused?
- Who wants to change page importance?

GOOGLE BOMBS



Google Web Images Video ^{New!} News Maps Desktop [more »](#)

miserable failure [Advanced Search](#)
[Preferences](#)

Web

[Why these results?](#)
www.google.com/googleblog These results may seem politically slanted. Here's what happened.

[President of the United States - George W. Bush](#)
Biography of the president from the official White House web site.
www.whitehouse.gov/president/ - 24k - 19 Sep 2006 - [Cached](#) - [Similar pages](#)


[Biography of Jimmy Carter](#)
Short biography from the official White House site.
www.whitehouse.gov/history/presidents/jc39.html - 31k - [Cached](#) - [Similar pages](#)

[BBC NEWS | Americas | 'Miserable failure' links to Bush](#)
Web users manipulate a popular search engine so an unflattering description leads to the president's page.
news.bbc.co.uk/2/hi/americas/3298443.stm - 32k - [Cached](#) - [Similar pages](#)

GOOGLE BOMBS



[Urban Dictionary: Worst Band In The World](#)  

Worst Band In The World - 2 definitions - "Creed" according to Google. Although it's fixed now, you used to be able to type in "the **worst band** in t...
www.urbandictionary.com/define.php?term=Worst%20Band%20In%20The%20World - 20k -
[Cached](#) - [Similar pages](#) - 

[The 50 Worst Artists in Music History Article on Blender :: The ...](#)    · [Music](#)

ARE YOU IN THE **WORST BAND IN THE WORLD**? Take this simple multiple-choice quiz and save yourself some embarrassment! 1 How long is your drummer's solo? ...
www.blender.com/guide/articles.aspx?id=466 - 140k - [Cached](#) - [Similar pages](#) - 

[Digg - Google search for "Worst band in the world" results in ...](#)  

When you Google **worst band in the world**, it brings up results for the **band** " Creed". Funny.
digg.com/software/Google_search_for_Worst_band_in_the_world_results_in_search_for_Creed - 54k - [Cached](#) - [Similar pages](#) - 

See results for: [creed](#)

[Creed.com – The Official Website of Creed](#)

Creed.com - the Official website of **Creed**. ... You can purchase and download ringtones of all your favorite **Creed** songs through our new mobile partner at ...
www.creed.com/

[Creed \(band\) - Wikipedia, the free encyclopedia](#)


Creed was an American post-grunge band from Tallahassee, Florida that became popular in the late 1990s and early 2000s. The band won a Grammy Award for Best ...
[en.wikipedia.org/wiki/Creed_\(band\)](http://en.wikipedia.org/wiki/Creed_(band))

GOOGLE BOMBS



murder

About 212,000,000 results (0.17 seconds) #

 Everything

 Images

 Videos

 News

 Shopping

 Realtime

 Books

 More

Houston, TX
Change location

▶ [Murder - Wikipedia, the free encyclopedia](#) ☆ 🔍

Murder is the unlawful killing of another human being with "malice aforethought", and generally this state of mind distinguishes **murder** from other forms of ...

[Murder \(United States law\)](#) - [Murder in English law](#) - [Murder \(Canadian law\)](#)
en.wikipedia.org/wiki/Murder - Cached - Similar

[Abortion - Wikipedia, the free encyclopedia](#) ☆ 🔍

Generally, the former position argues that a human fetus is a human being ...

[United States](#) - [Methods of abortion](#) - [Abortion by country](#) - [Abortion law](#)
en.wikipedia.org/wiki/Abortion - Cached - Similar

[+](#) Show more results from wikipedia.org

[Murder | Define Murder at Dictionary.com](#) ☆ 🔍

Law. the killing of another human being under conditions specifically covered in law. In the U.S., special statutory definitions include **murder** committed ...

dictionary.reference.com/browse/murder - Cached - Similar

BEYOND GOOGLE BOMBS

- Search engine optimization is a major area
- Arms race between google and web companies
- At equilibrium(?)



HARDWARE



IMAGE SEARCH HISTORY

- Jennifer Lopez wore this dress to the 2000 Grammys
- Most popular Google search ever; Google had no way to obtain images
- Google Image Search is born



IMAGE SEARCH

- How does that work??

[paraphrasing]
Computers can't
really analyze
what's in an image

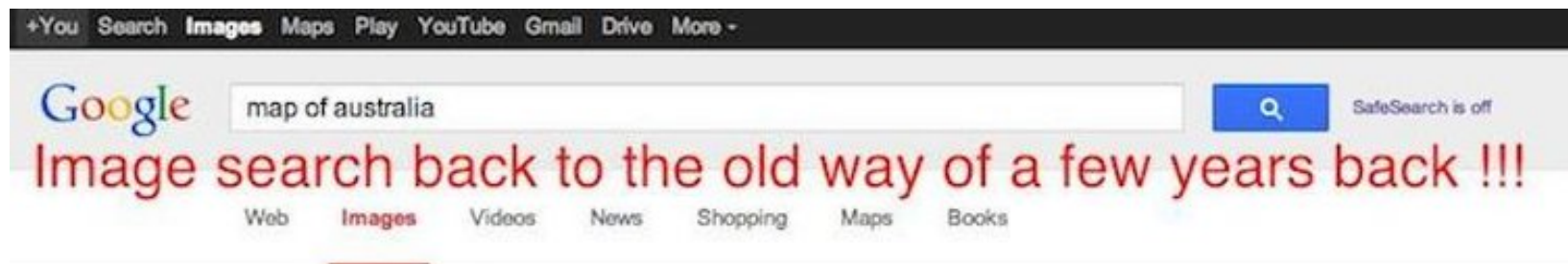


IMAGE SEARCH ORIGINALLY

- “Context based” (text-based, concept-based)
- Look around the picture for words, etc.
- Look in metadata

- When does this work well?

IMAGE SEARCH ORIGINALLY



About 113,000,000 results

Any size

Large
Medium
Icon

Any color

Full color
Black and white
Transparent

Any type

Face
Photo
Clip art
Line drawing
Animated

Any time

Past 24 hours
Past week



lonelyplanet.com
Map of Australia
466 × 350 - 64k - jpg



australia.edu
Printable Map of Australia
1424 × 1270 - 1703k - jpg



maps of world.com
Map of Australia
1000 × 854 - 228k - gif



au.totaltravel.yahoo.com
Map of Australia
638 × 650 - 53k - gif



en.wikipedia.org
Australia map
2190 × 1577 - 134k - png



ezilon.com
Australia Political Map
1391 × 998 - 353k - gif



sydney-australia.biz
Map of Australia - Click to
940 × 709 - 161k - png



ritas-outback-guide.com
The Australian map pictured
620 × 620 - 85k - jpg

IMAGE SEARCH ORIGINALLY

The screenshot shows a web browser window titled "nature - Google Search". The address bar contains the URL <http://www.google.com/images?hl=en&source=imghp&biw=1004&bih=609&q>. The search bar contains the text "nature" and a "Search" button. Below the search bar, it says "About 420,000,000 results (0.05 seconds)" and "Advanced search".

On the left side, there is a navigation menu with "Everything", "Books", "Images" (selected), and "More". Below this, there are filter options for "Any size" (Medium, Large, Icon, Larger than..., Exactly...), "Any type" (Face, Photo, Clip art, Line drawing), and "Any color" (Full color, Black and white, with a color palette).

The main content area displays a grid of image search results. Each result includes a small image, a title, dimensions, file size, and source URL, along with a "Find similar images" link. The results shown are:

- Speak Your Mind**: 500 x 375 - 52k - jpg, underprocess.com, Find similar images
- Unplug that Plasma Screen,**: 590 x 472 - 129k - jpg, psychologytoday.com, Find similar images
- Nature Photos**: 1024 x 768 - 259k - jpg, interweb.in, Find similar images
- nature**: 1600 x 1200 - 566k - jpg, musikality.net, Find similar images
- Riverside Nature**: 2048 x 1536 - 728k - jpg, kervilletexascvb.com, Find similar images
- Nature is amazing**: 503 x 516 - 319k - gif, people.ucsc.edu, Find similar images
- Nature Spirits ~**: 556 x 350 - 32k - jpg, crystalinks.com, Find similar images
- Nature Speaks**: 800 x 600 - 53k - jpg, called2ministry..., Find similar images

IMAGE SEARCH NOW

- Can't be context-based: reverse image search
- <https://www.google.com/imghp?hl=en>
- Facebook “may contain”
- For accessibility

HOW DOES IT WORK?

- Machine learning!
- We'll discuss tomorrow

VIDEO SEARCH?

- Still (mostly?) context-based
- No reverse video search
- Youtube search isn't that good

WHAT DOESN'T GET INDEXED?



DEEP WEB

- All of the web that can't be indexed by search engines
- Includes wayback machine, lots of documents/videos
- Williams databases
- Private Facebook
- Anything query-based
- Very big!

INDEXING THE DEEP WEB?

Crawling the Hidden Web

Sriram Raghavan, Hector Garcia-Molina
Computer Science Department, Stanford University
Stanford, CA 94305, USA
{rsram, hector}@cs.stanford.edu

Abstract

Current-day crawlers retrieve content only from the publicly indexable Web, i.e., the set of web pages reachable purely by following hypertext links, ignoring search forms and pages that require authorization or prior registration. In particular, they ignore the tremendous amount of high quality content “hidden” behind search forms, in large searchable electronic databases. In this paper, we provide a framework for addressing the problem of extracting content from this hidden Web. At Stanford, we have built a task-specific hidden Web crawler called the Hidden Web Exposer (HiWE). We describe the architecture of HiWE and present a number of novel techniques that went into its design and implementation. We also present results from experiments we conducted to test and validate our techniques.

Keywords: Crawling, Hidden Web, Content extraction, HTML Forms

1 Introduction

A number of recent studies [4, 19, 20] have noted that a tremendous amount of content on the Web is *dynamic*. This dynamism takes a number of different forms (see Section 2). For instance, web pages can be dynamically generated, i.e., a server-side program creates a page *after* the request for the page is received from a client. Similarly, pages can be dynamic because they include code that executes on the client machine to retrieve content from remote servers (e.g., a page with an embedded applet that retrieves and displays the latest stock information).

Based on studies conducted in 1997, Lawrence and Giles [19] estimated that close to 80% of the content on the Web is dynamically generated, and that this number is continuing to increase. As major software vendors come up with new technologies [2, 17, 26] to make such dynamic page generation simpler and more efficient, this trend is

HOW TO FILL OUT WEB FORMS?

- Look at how other people have filled out similar forms
- Guess and repeat

SITEMAPS

- Protocol to allow server to tell crawler about web pages
- Don't need to follow links! Just get a list



DARK WEB

- Needs special software to access
 - Not just a web browser
- Tor, etc.
- Largely used for illegal activity

DARK WEB

KIM ZETTER SECURITY 02.10.15 10:17 AM

Darpa Is Developing a Search Engine for the Dark Web

A NEW SEARCH engine being developed by Darpa aims to shine a light on the dark web and uncover patterns and relationships in online data to help law enforcement and others track illegal activity.

The project, dubbed Memex, has been in the works for a year and is being developed by 17 different contractor teams who are working with the military's Defense Advanced Research Projects Agency. Google and Bing, with search results influenced by popularity and ranking, are only able to capture approximately five percent of the internet. The goal of Memex is to build a better map of more internet content.

"The main issue we're trying to address is the one-size-fits-all

TOR

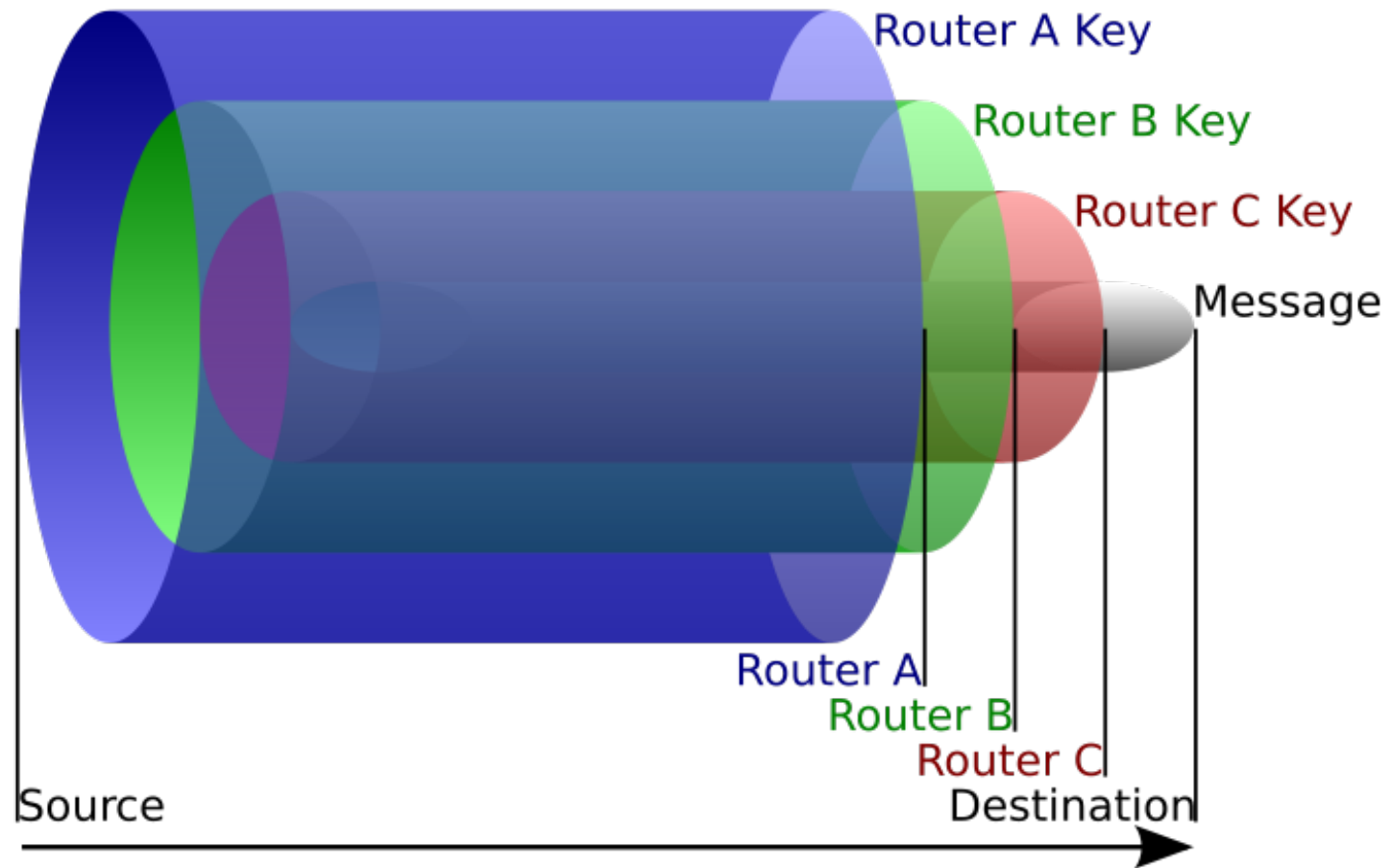
- Reminder: who can see what site you're communicating with?
- What does HTTPS do?
- Isn't this impossible to avoid?
 - Can't send a letter without an address, right?

TOR

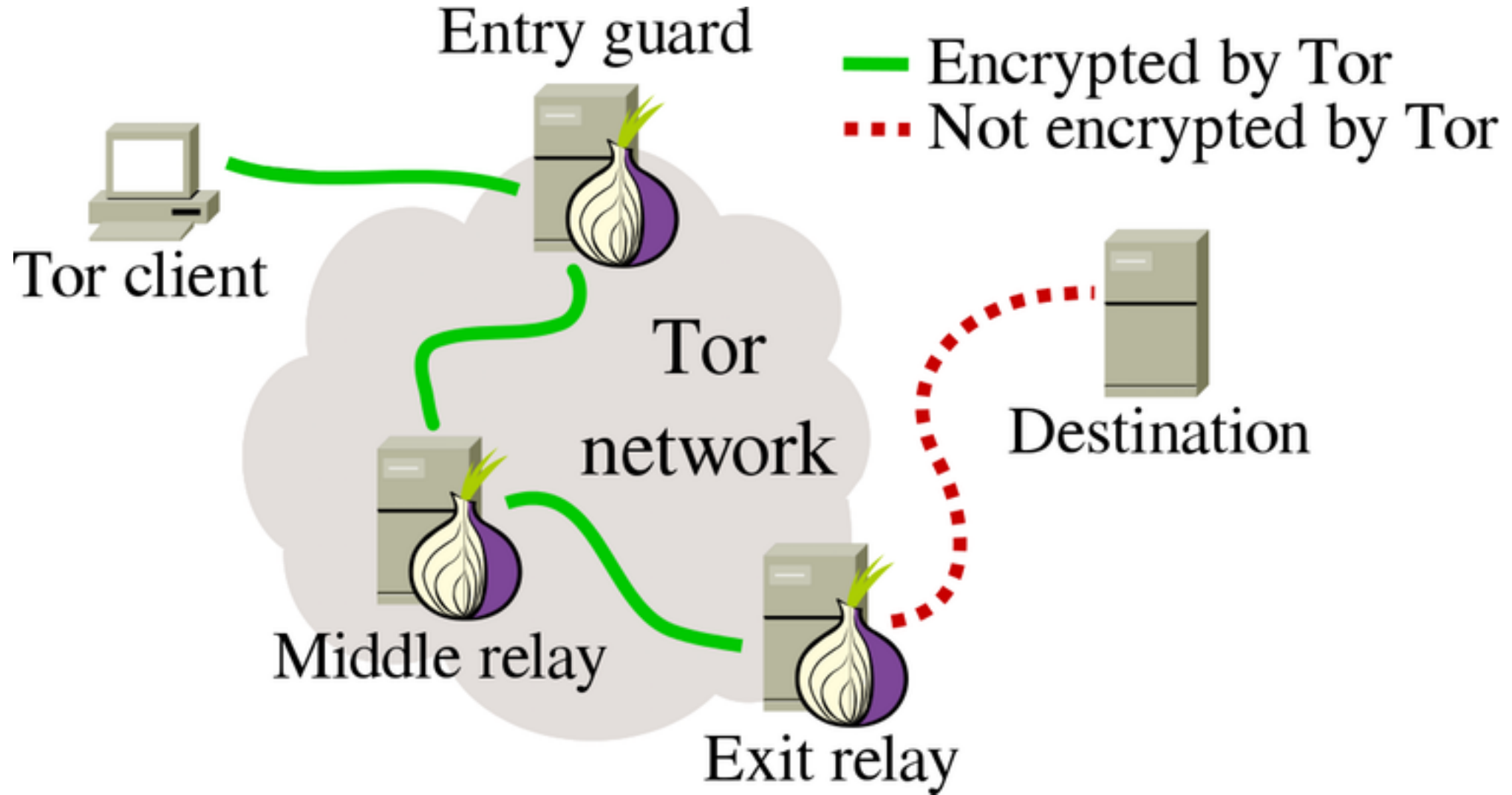
- “The onion router”



TOR



TOR

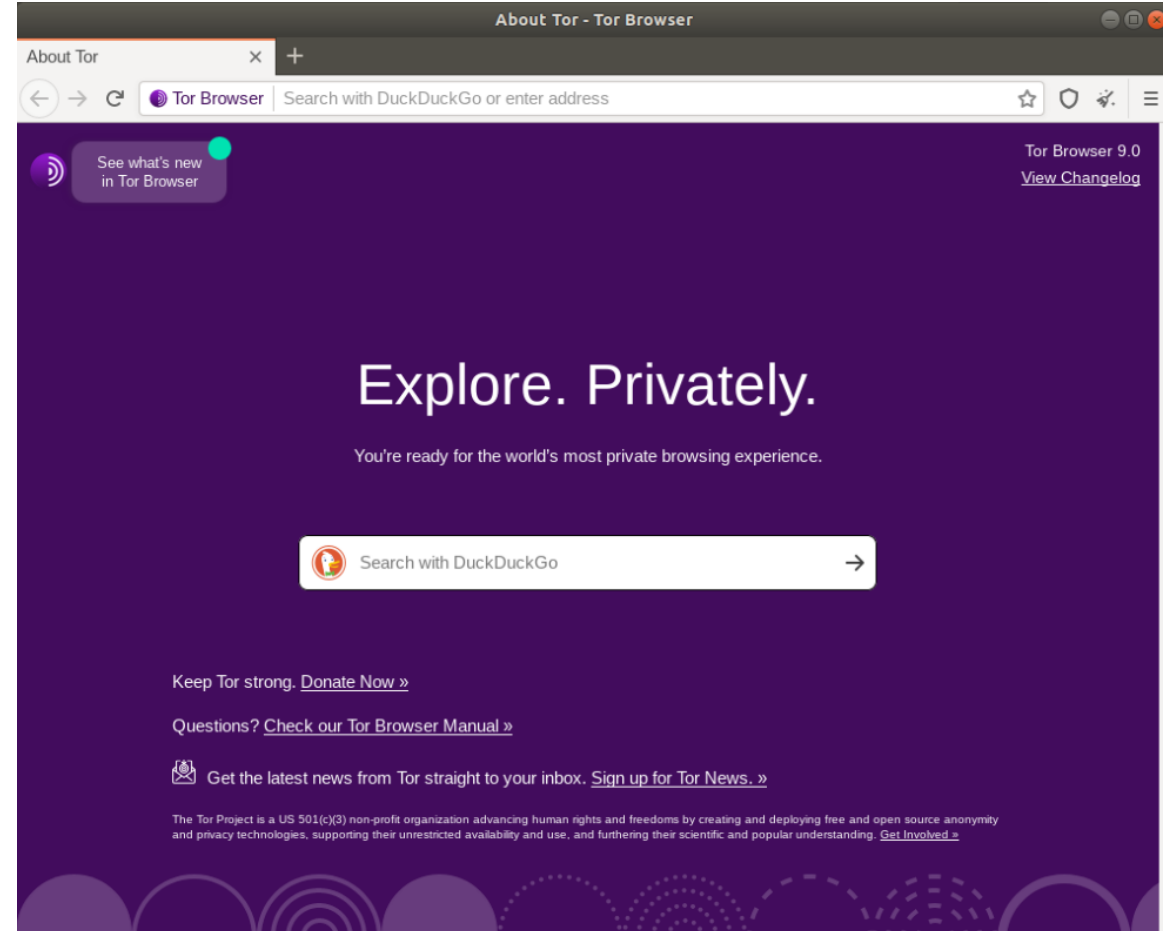


TOR

- Each server only knows:
 - Last server you contacted
 - Next server you want to contact
- None knows the whole route (assuming encryption works)

TOR: HOW TO USE

- Browser
- Many security-aware sites
- Why not use tor all the time?



TOR: IS IT SECURE?

- **Timing attack**
 - Look at when Nancy is accessing the internet, see what sites are accessed
- **Exit node attack**
 - Look for data leaving the tor network
- **Fingerprinting (like we discussed)**
 - Tor browser is pretty secure, but more difficult fingerprints like mouse movements can work

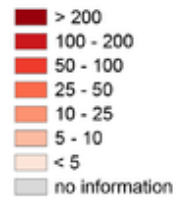
TOR: IS IT SECURE?

- If someone wants to deanonymize *you*, they probably can
- If someone wants to deanonymize *everyone*, they probably can't
- ??

TOR USAGE

The anonymous Internet

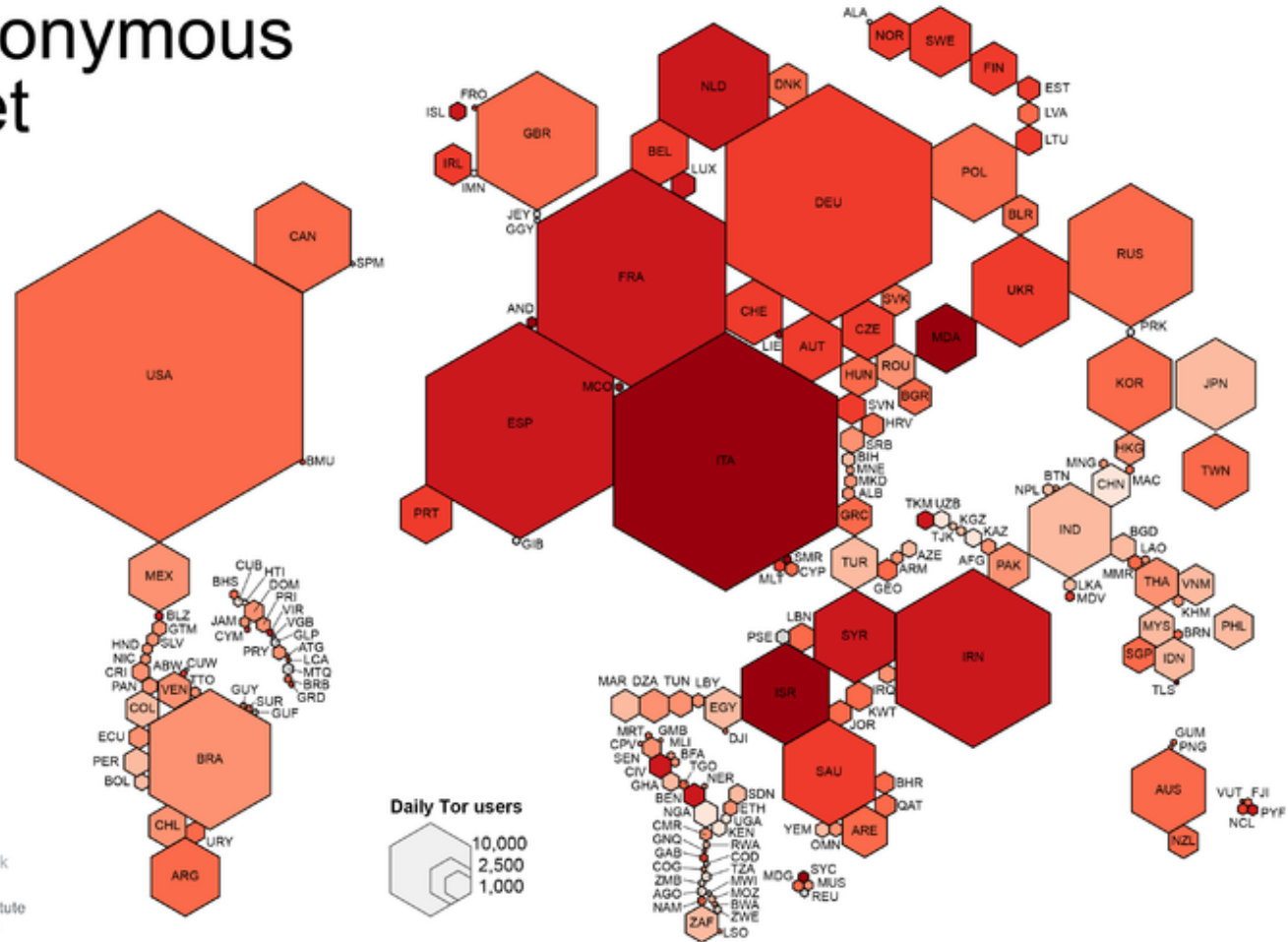
Daily Tor users
per 100,000
Internet users



Average number of
Tor users per day
calculated between
August 2012 and
July 2013

data sources:
Tor Metrics Portal
metrics.torproject.org
World Bank
data.worldbank.org

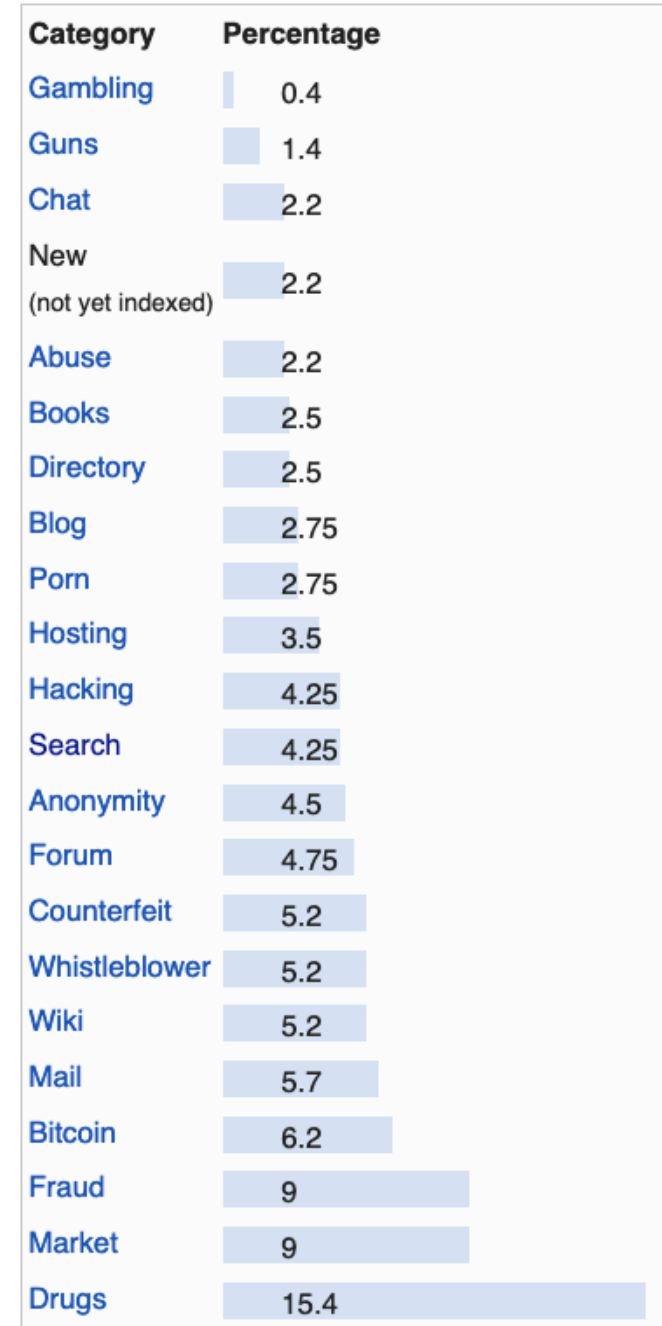
by Mark Graham
(@geoplace) and
Stefano De Sabbata
(@maps4thought)
Internet Geographies at
the Oxford Internet Institute
2014 • geography.oii.ox.ac.uk



TOR USAGE

Rule of thumb: ~20-40%
legal/legitimate usage

Web-based onion services in January 2015^[23]



SILK ROAD

- About \$15 million/year in traffic
- Largely drugs
- “Unregulated market”



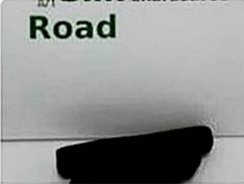









Silk Road
anonymous market

messages 1 | orders 0 | account ₪0.00

Search Go

Shop by Category

- Drugs 2,399
 - Cannabis 341
 - Dissociatives 65
 - Ecstasy 209
 - Opioids 156
 - Other 144
 - Precursors 12
 - Prescription 526
 - Psychedelics 427
 - Stimulants 273
- Apparel 114
- Art 7
- Books 743
- Collectibles 12
- Computer equipment 19
- Custom Orders 26
- Digital goods 310
- Drug paraphernalia 89
- Electronics 20
- Erotica 319
- Fireworks 2
- Food 3
- Forgeries 58
- Hardware 2
- Home & Garden 7
- Jewelry 48
- Lab Supplies 5
- Lotteries & games 29
- Medical 5

 <p>5x - 10mg Dexedrine (Pure Dextroamphetamine) ₪4.94</p>	 <p>2 x 0,25 mg Xanax (Alprazolam) ₪1.50</p>	 <p>Malana charas hand rubbed Indian hash 100g ₪75.83</p>	 <p>1 Gram OG KUSH OIL 81% THC 90% TOTAL ₪4.13</p>
 <p>14 grams (1/2 Ounce) of Nebula JWH-122 ₪2.63</p>	 <p>3.5g Crystal Meth Ice Shards ₪31.92</p>	 <p>20 x 25mg Cialis ₪2.57</p>	 <p>!!!...Psilocybe-Cubensis-Chocolate...!!! ₪18.15</p>
 <p>100 x Orange Star Very high MDMA content 180mg</p>	 <p>100x 200mg White XTC 'Speakers'</p>	 <p>3g Methylone Crystals -₪50-Lab Grade</p>	 <p>15mg Adderall Extended Release (1 Capsule)</p>

SILK ROAD

Shut down in October 2013
Creator Ross Ulbricht arrested



THIS HIDDEN SITE HAS BEEN SEIZED

by the Federal Bureau of Investigation,
in conjunction with the IRS Criminal Investigation Division,
ICE Homeland Security Investigations, and the Drug Enforcement Administration,
in accordance with a seizure warrant obtained by the
United States Attorney's Office for the Southern District of New York
and issued pursuant to 18 U.S.C. § 983(j) by the
United States District Court for the Southern District of New York



SILK ROAD SEIZURE



HOW WAS THE SITE FOUND?

- Hosted in Iceland
- FBI claims it was found via data from the site
CAPTCHA

HOW WAS THE CREATOR FOUND?

- IRS agent inspired by Al Capone takedown



FINDING THE CREATOR

- Gary Alford used google to find the first mention of silk road he could...




altoid
Stranger
Registered: 01/27/11
Posts: 1
Last seen: 6 years, 7 months

anonymous market online? **NEW** 29
#13860995 - 01/27/11 11:28 PM (6 years, 7 months ago)

I came across this website called Silk Road. It's a Tor hidden service that claims to allow you to buy and sell anything online anonymously. I'm thinking of buying off it, but wanted to see if anyone here had heard of it and could recommend it.

I found it through silkroad420.wordpress.com, which, if you have a tor browser, directs you to the real site at <http://tydgccykixpbu6uz.onion>.

Let me know what you think...

Post Extras:   

LOOKING FOR “ALTOID”

- From bitcoin talk
- Deleted, but bits remained

Quote from: altoid on January 29, 2011, 07:44:51 PM

What an awesome thread! You guys have a ton of great ideas. Has anyone seen Silk Road yet? It's kind of like an anonymous amazon.com. I don't think they have heroin on there, but they are selling other stuff. They basically use bitcoin and tor to broker anonymous transactions. It's at <http://tydgccykixpbu6uz.onion>. Those not familiar with Tor can go to silkroad420.wordpress.com for instructions on how to access the .onion site.

SMOCKING GUN



Author

Topic: IT pro needed for venture backed bitcoin startup (Read 36386 times)

altoid

Jr. Member



Activity: 48



IT pro needed for venture backed bitcoin startup

October 11, 2011, 08:06:22 PM

#1

Hello, sorry if there is another thread for this kind of post, but I couldn't find one. I'm looking for the best and brightest IT pro in the bitcoin community to be the lead developer in a venture backed bitcoin startup company. The ideal candidate would have at least several years of web application development experience, having built applications from the ground up. A solid understanding of oop and software architecture is a must. Experience in a start-up environment is a plus, or just being super hard working, self-motivated, and creative.

Compensation can be in the form of equity or a salary, or somewhere in-between.

If interested, please send your answers to the following questions to [rossulbricht at gmail dot com](mailto:rossulbricht@gmail.com)

- 1) What are your qualifications for this position?
- 2) What interests you about bitcoin?

From there, we can talk about things like compensation and references and I can answer your questions as well. Thanks in advance to any interested parties. If anyone knows another good place to recruit, I am all ears.

OTHER EVIDENCE

- Early admin sign-on from nearby suspect's address
- Stackoverflow questions
 - Username "frosty" matched computer name
- Seized package with 9 fake IDs headed to his address
 - At the time, told agents "anyone can go on a site called Silk Road and buy fake IDs"

FINAL ARREST

- <https://www.businessinsider.com/the-arrest-of-silk-road-mastermind-ross-ulbricht-2015-1>