The Web IR - Challenges

Vasudeva Varma

IIIT Hyderabad

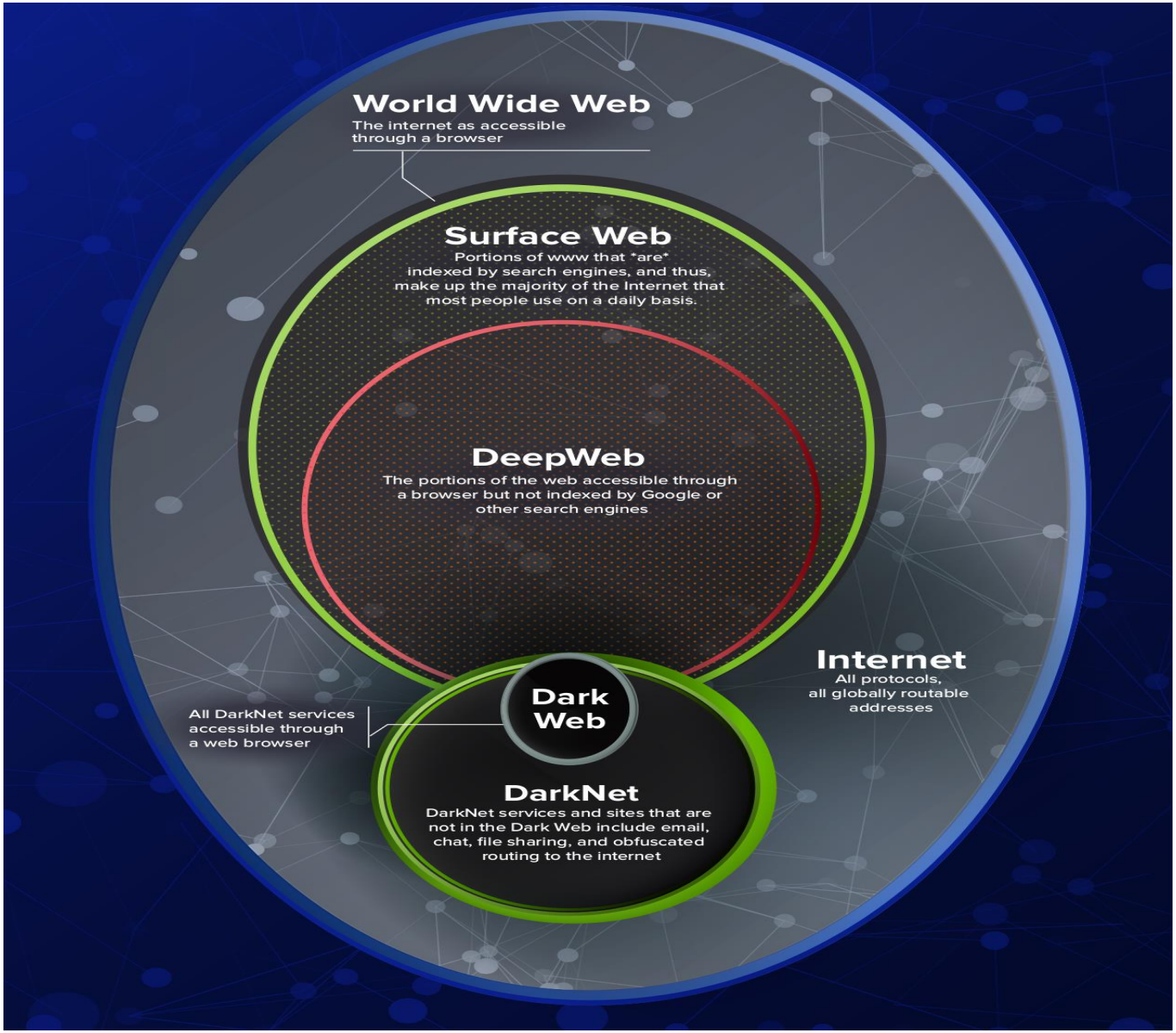
The web and its challenges

- Unusual and diverse documents
- Unusual and diverse users, queries, information needs
- Beyond terms, exploit ideas from social networks
  - link analysis, clickstreams ...


Surface Web
Deep Web
Dark Web

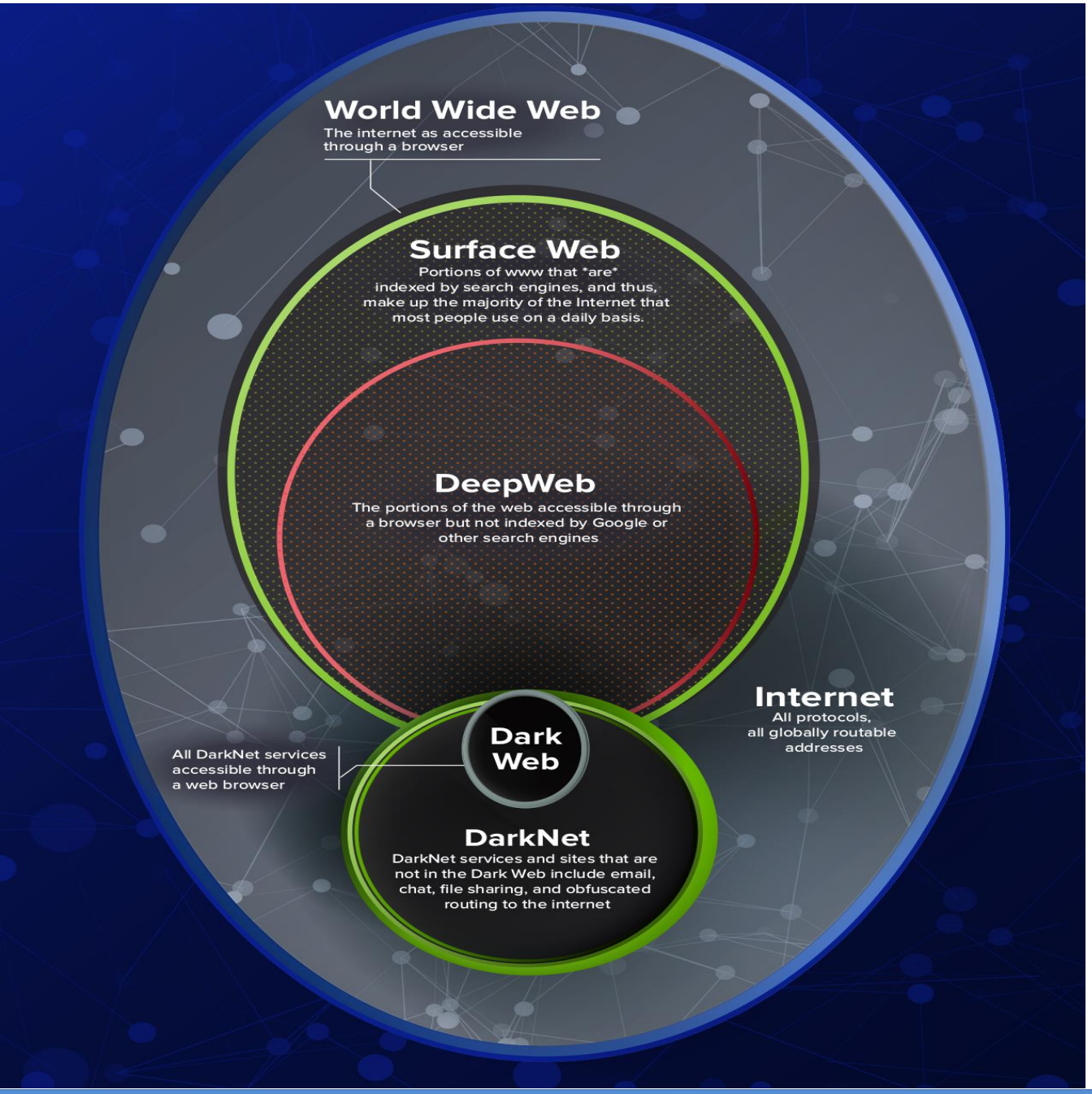

SURFACE WEB
Google
Bing
Wikipedia
DEEP WEB
Academic Information
Medical Records
Legal Documents
Scientific Reports
Subscription Information
Contains 90% of the information on the Internet, but is not accessible by Surface Web crawlers.
Social Media
Multilingual Databases
Financial Records
Government Resources
Competitor Websites
Organization-specific Repositories
(DARK WEB)
A part of the Deep Web accessible only through certain browsers such as Tor designed to ensure anonymity. Deep Web Technologies has zero involvement with the Dark Web.
Illegal Information
TOR-Encrypted sites
Political Protests
Drug Trafficking sites
Private Communications

Surface web, deep web, dark web…


World Wide Web
The internet as accessible through a browser
Surface Web
Portions of www that *are* indexed by search engines, and thus, make up the majority of the Internet that most people use on a daily basis.
DeepWeb
The portions of the web accessible through a browser but not indexed by Google or other search engines
Internet
All protocols, all globally routable addresses
All DarkNet services accessible through a web browser
Dark Web
DarkNet
DarkNet services and sites that are not in the Dark Web include email, chat, file sharing, and obfuscated routing to the internet

The web: size

- What is being measured?
  - Number of hosts
  - Number of (static) html pages
    - Volume of data
- Number of hosts – netcraft survey
  - http://news.netcraft.com/archives/web_server_survey.html
  - Gives monthly report on how many web servers are out there
- Number of pages – numerous estimates
  - For a Web engine: how big its index is
- https://www.internetlivestats.com/



What is the size of DEEP web?

Crawling…

Crawling overview

- Types of crawlers
- Functionality
  - Start the crawl: Seed URLs
  - …
  - End the crawl: Halting Criteria
- Policies
- Architecture of a Web crawler

Types of crawlers

- Search engine crawlers
- Enterprise crawlers
- Monitoring crawlers
  - Copy right violations
  - DRM crawlers
  - Malware detection
  - Web analytics
- Document feeds (RSS/Atom or commercial feeds)

Functionality of the crawlers

- Start with seed URLs
  - Selection of seed URLs is important
    - Quality (avoid spam/objectionable/non-hub pages)
    - Importance (popularity/trustworthiness/reliability)
    - Potential yield documents
  - Web graph helps pick right seed URLs
- Survive
  - Avoid crawler traps
    - Causes infinite number of requests being made
    - Infinitely deep directory structures
  - Follow the rules and behave well (adhere to ***policies***)
- End when time comes
  - Some crawlers are designed to go on forever
  - Some stop when a particular criteria is met (after reaching depth K, after crawling N pages or time T, after Index reaches K Units)

Policies

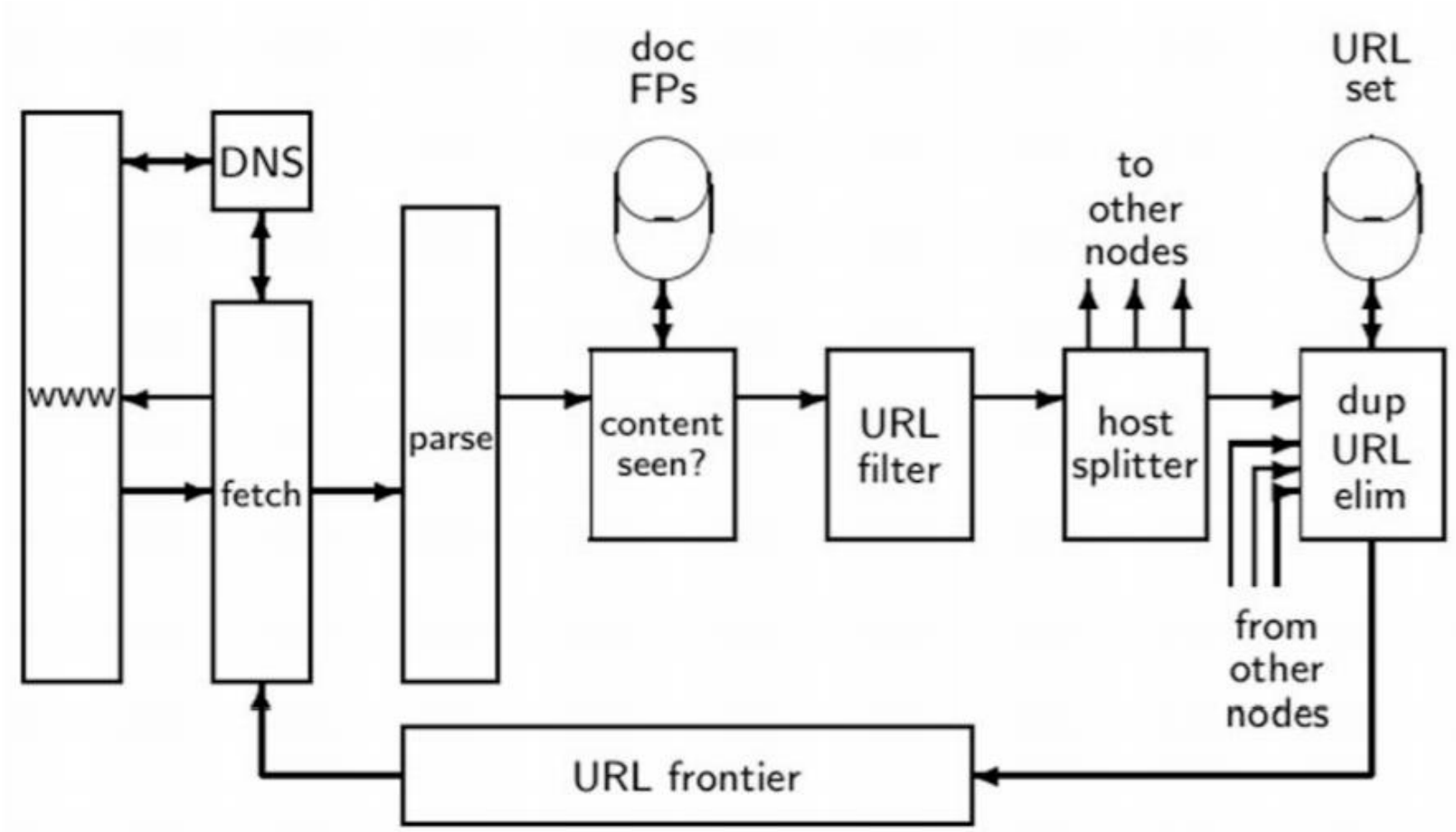
The behavior of a Web crawler is the outcome of a combination of policies

- Selection policy: states which pages to download
  - Prioritization: predict high yield pages from web graph
- Revisit policy: states when to check for changes to the pages
  - goal: high avg Freshness and low avg age
  - Two policies: Uniform policy or proportional policy
- Politeness policy: states how to avoid overloading Web sites
  - Robots.txt
  - Sitemap – organize the site to control crawling its parts
  - Meta tag: `<META NAME “ROBOTS” CONTENT=“NOINDEX, NOFOLLOW”>`
- Parallelization policy: states how to coordinate distributed Web crawlers

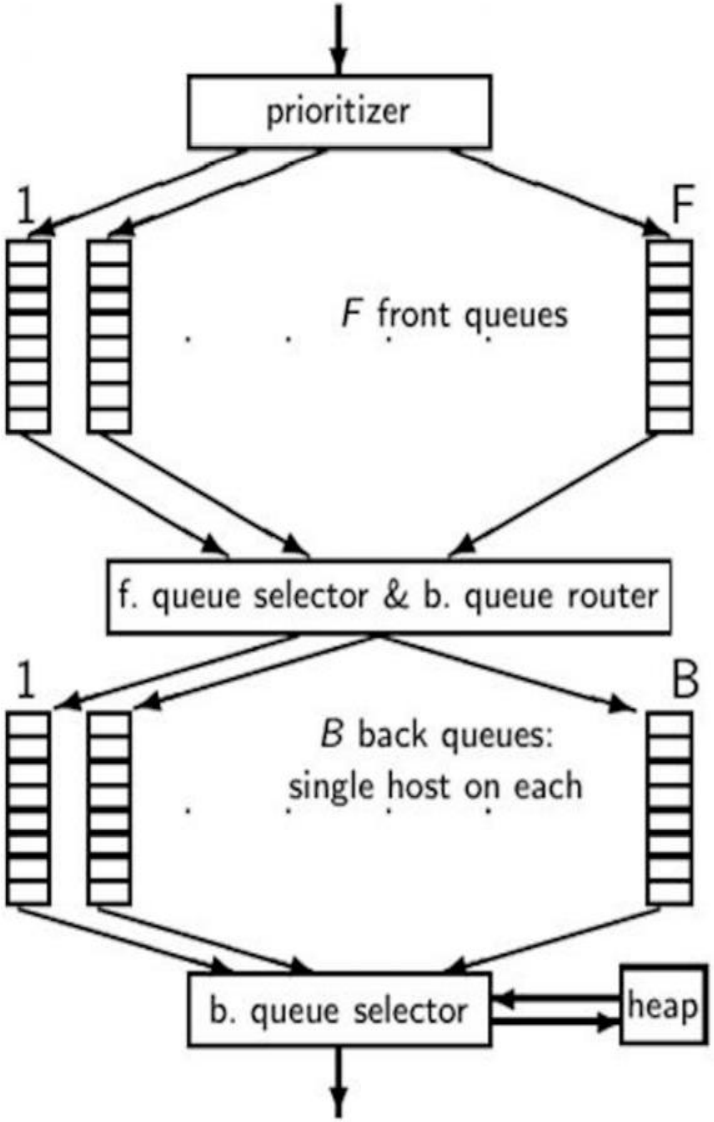
Challenge: How to handle politeness and priority simultaneously?

- An ideal crawler needs to frequently visit high priority pages but also needs to be polite.
- To achieve this we implement a URL frontier with the following goals.
  - only one connection is open at a time to any host
  - a waiting time of a few seconds occurs between successive requests to a host
  - high-priority pages are crawled preferentially

Distributed Crawler Architecture



URL Frontier Design



Examples of Popular Web Crawlers

- GooleBot
- BingBot
- MSNbot
- Slug
- Yahoo!Slurp

- Nutch
- Scrapy
- DataParkSearch
- Grub
- Heritrix

Further reading

- Introductory article on deep web link
- Section on crawling the deep web, from this university's guide on deep web:
- Optional Reading: Paper on crawling the deep web
- Victor Lavrenko short videos on web crawling
- Implementation view