Jul – Dec 2022
IT458

Information Retrieval on the Web

Link Analysis and Web Search

A Minute on the Internet in 2019

Estimated data created on the internet in one minute

Sources: Lori Lewis & Officially Chad via Visual Capitalist

A Minute on the Internet in 2020

Estimated amount of data created on the internet in one minute

NETFLIX
404,444 hours of video streamed by users

TikTok
2,704 app installations

amazon
6,659 packages shipped

zoom
208,333 participants in meetings

319 new users gained

500 hours of video uploaded by users

347,222 stories

52,083 users connected

41.7m messages shared

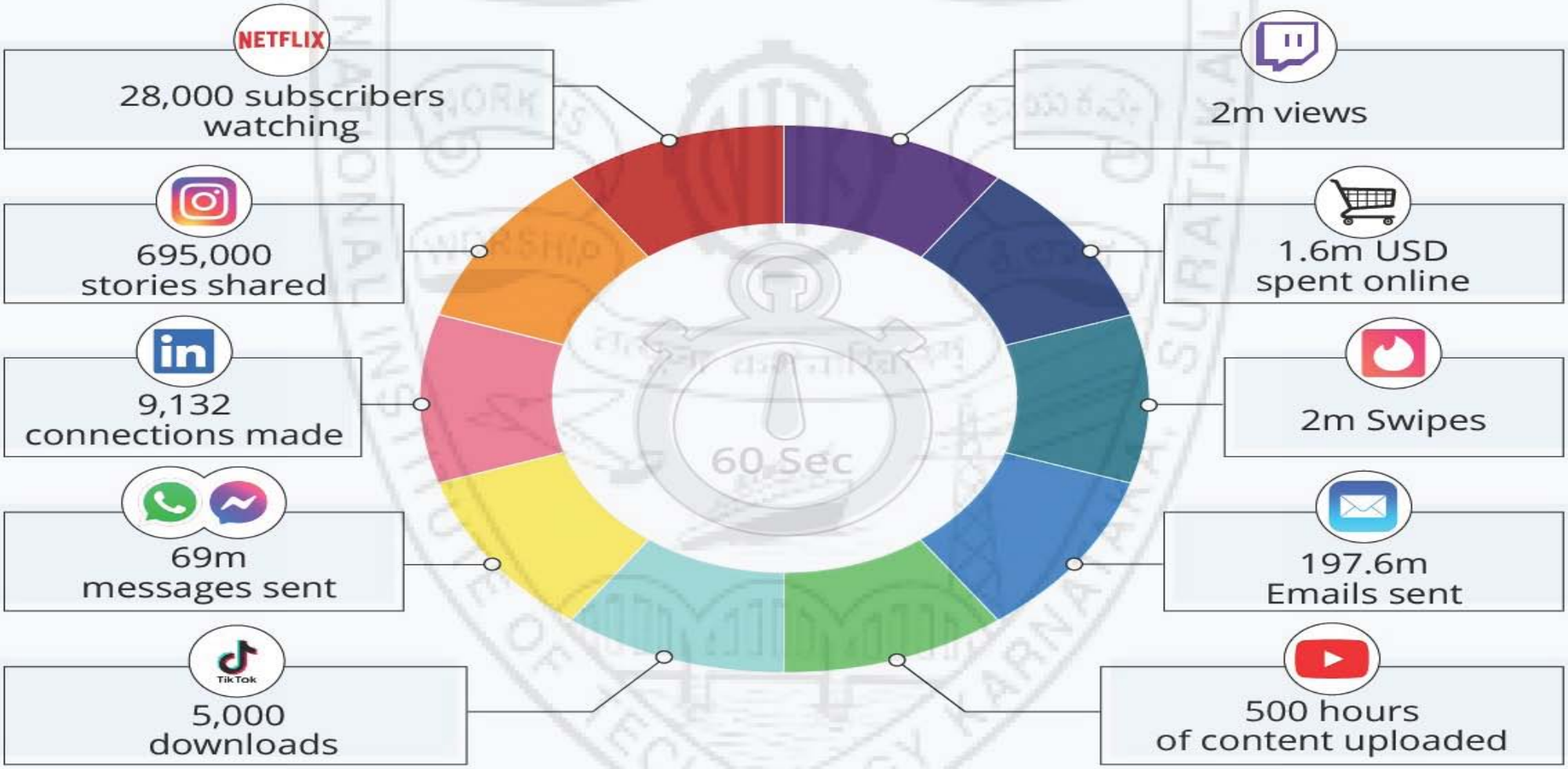
28 new tracks added to library

Source: Visual Capitalist

statista

A Minute on the Internet in 2021

Estimated amount of data created on the internet in one minute



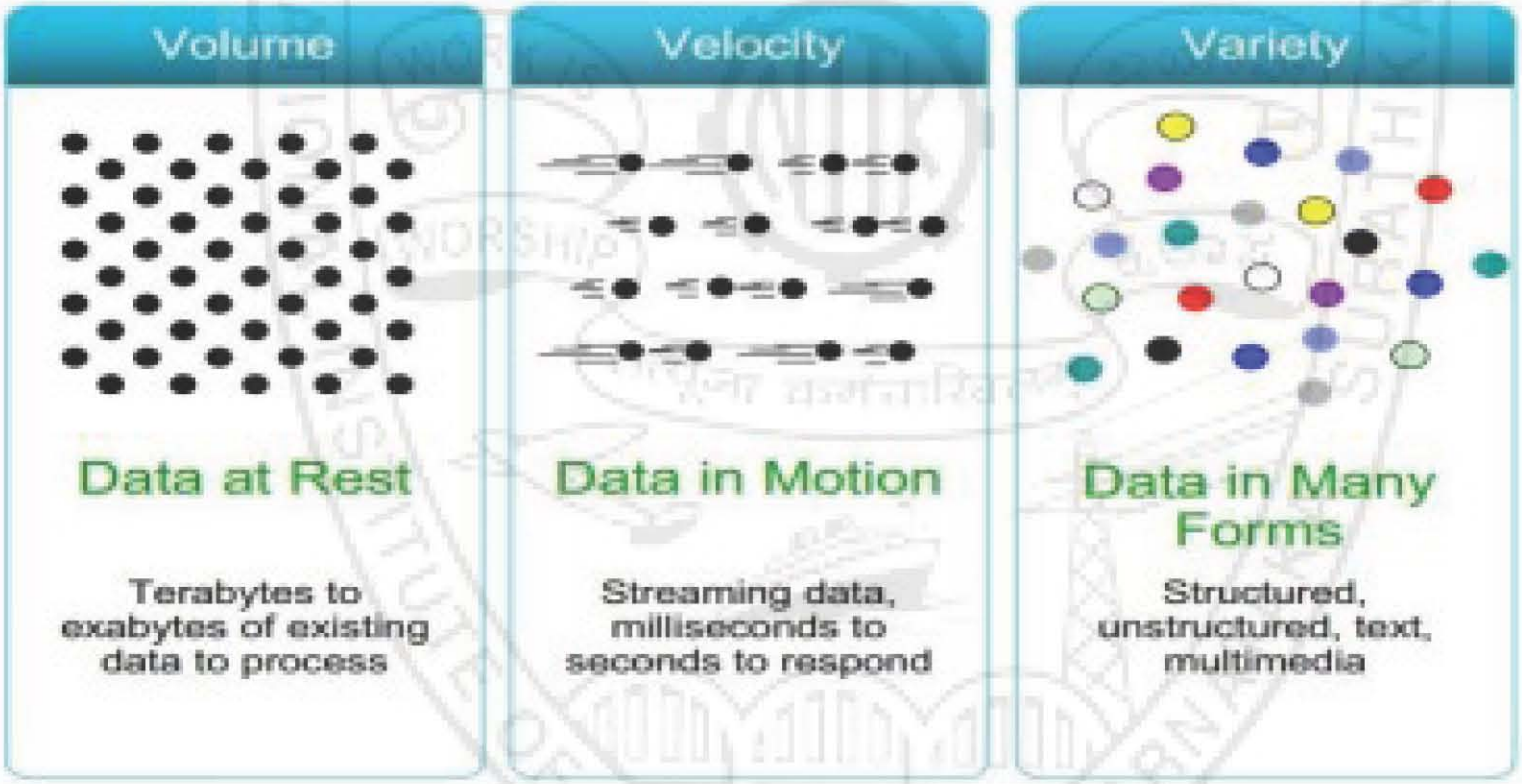
Source: Lori Lewis via AllAccess


2020 This Is What Happens In An Internet Minute
facebook
1.3 Million Logging In
19 Million Texts Sent
YouTube
4.7 Million Videos Viewed
400,000 Apps Downloaded
694,444 Scrolling Instagram
194,444 People Tweeting
1.6 Million Swipes
tinder
190 Million Emails Sent
1.2 Million Views
twitch
1,400 Downloads
Tik Tok
305 Smart Speakers Shipped
amazon echo
2.5 Million Images Viewed
imgur
59 Million Messages Sent
2.5 Million Snaps Created
$1.1 Million Spent Online
NETFLIX
764,000 Hours Watched
Google
4.1 Million Search Queries
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd
2021 This Is What Happens In An Internet Minute
facebook
1.4 Million Scrolling
LinkedIn
9,132 Connections Made
21.1 Million Texts Sent
YouTube
500 Hours Content Uploaded
414,764 Apps Downloaded
695,000 Stories Shared
200,000 People Tweeting
2 Million Swipes
tinder
197.6 Million Emails Sent
2 Million Views
twitch
5,000 Downloads
Tik Tok
932 Smart Audio Devices Shipped
amazon echo
3 Million Images Viewed
imgur
69 Million Messages Sent
3.4 Million Snaps Created
$1.6 Million Spent Online
NETFLIX
28,000 Subscribers Watching
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd

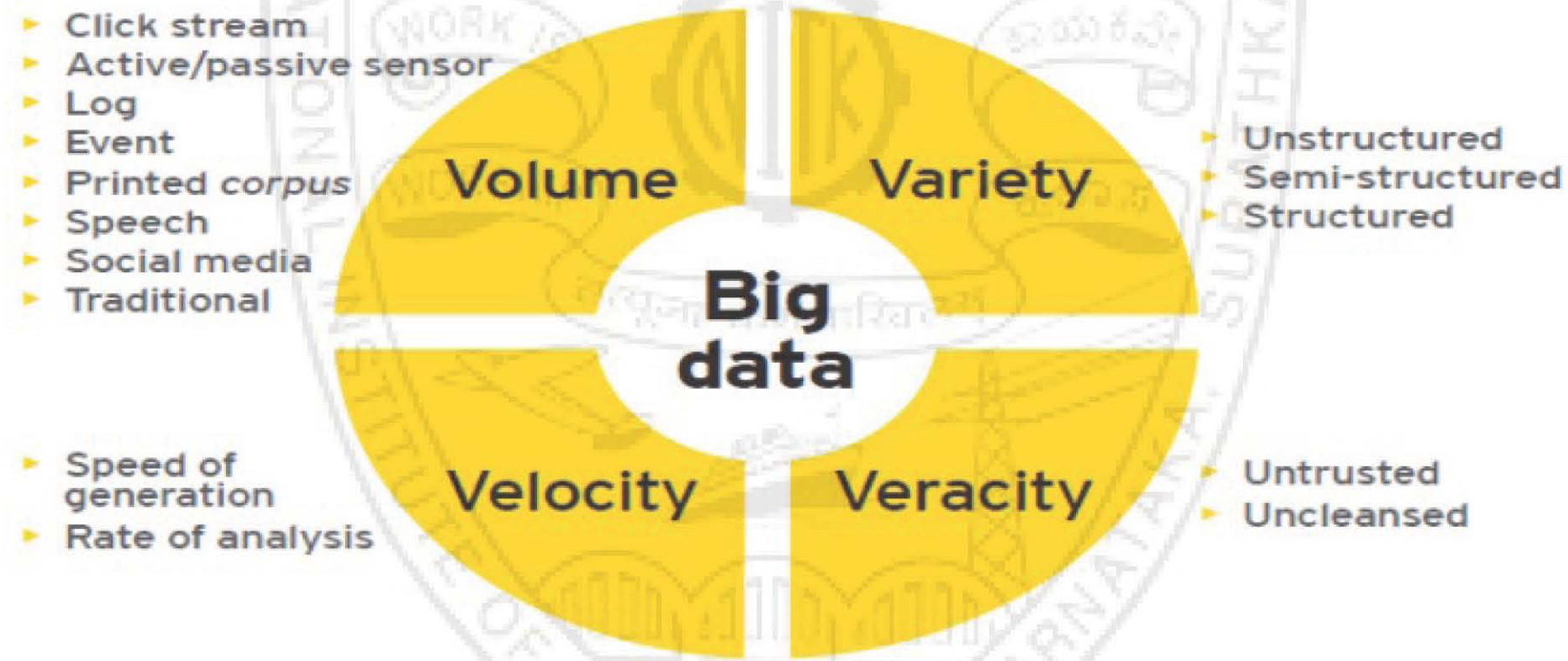
Web Search

- Main challenges in searching on the Web -
  - **Data-centric:** *related to data itself*
  - **Interaction-centric:** *related to users and their interactions*

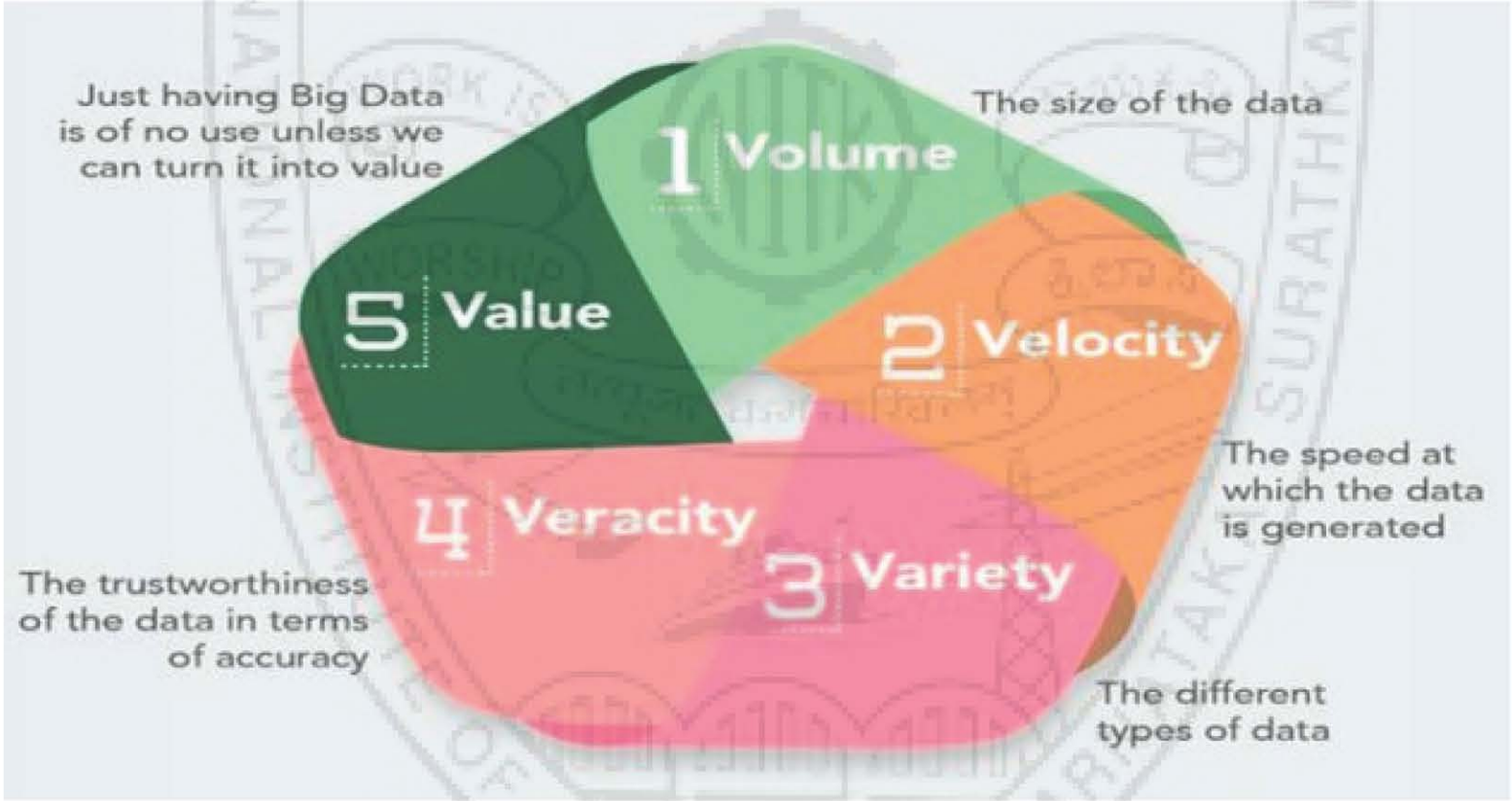
Web Search: Data centric Challenges



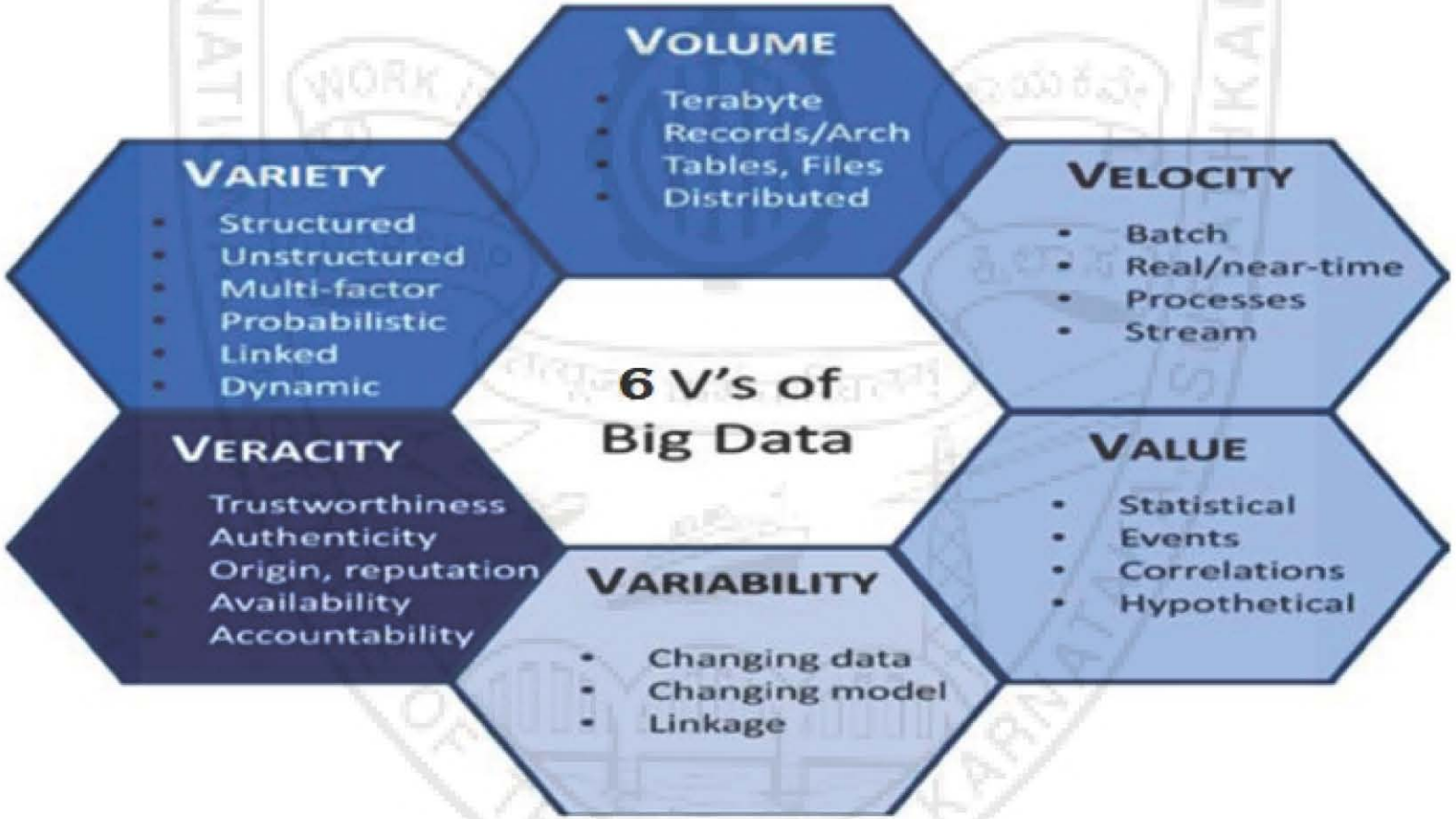
Web Search: Data centric Challenges



Web Search: Data centric Challenges



Web Search: Data centric Challenges



Web Search: Data centric Challenges

Web Search: Data centric Challenges

Web Search

- Main challenges in searching on the Web -
  - **Data-centric:** *related to data itself*
    - distributed data
    - volatile data
    - large volume of data
    - heterogeneous data
    - unstructured and redundant data
    - quality of data
    - Domain-specific nature of data
    - Value of data for varied applications
  - **Interaction-centric:** *related to users and their interactions*

Web Search

- Main challenges in searching on the Web -
  - **Data-centric:** *related to data itself*
    - distributed data
    - volatile data
    - large volume of data
    - heterogeneous data
    - unstructured and redundant data
    - quality of data
    - Domain-specific nature of data
    - Value of data for varied applications
  - **Interaction-centric:** *related to users and their interactions*
    - expressing a query
    - interpreting results

WorldWideWebSize.com

DAILY ESTIMATED SIZE OF THE WORLD WIDE WEB

The size of the World Wide Web (The Internet)

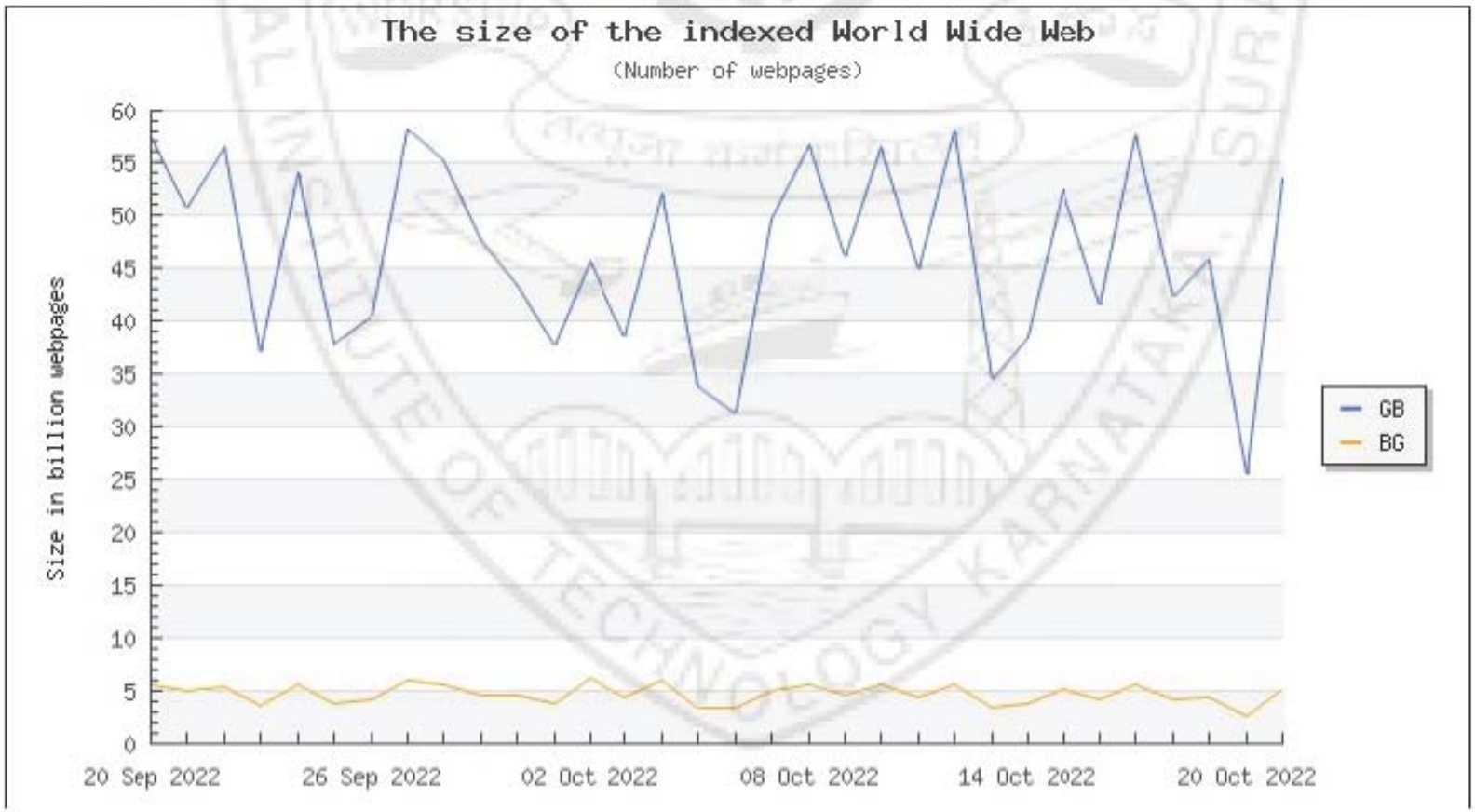
133 Like Share

The Indexed Web contains **at least 5.29 billion pages** (Friday, 21 October, 2022).

The Dutch Indexed Web contains **at least 2606.42 million pages** (Friday, 21 October, 2022).

The Indexed Web | The Dutch Indexed Web

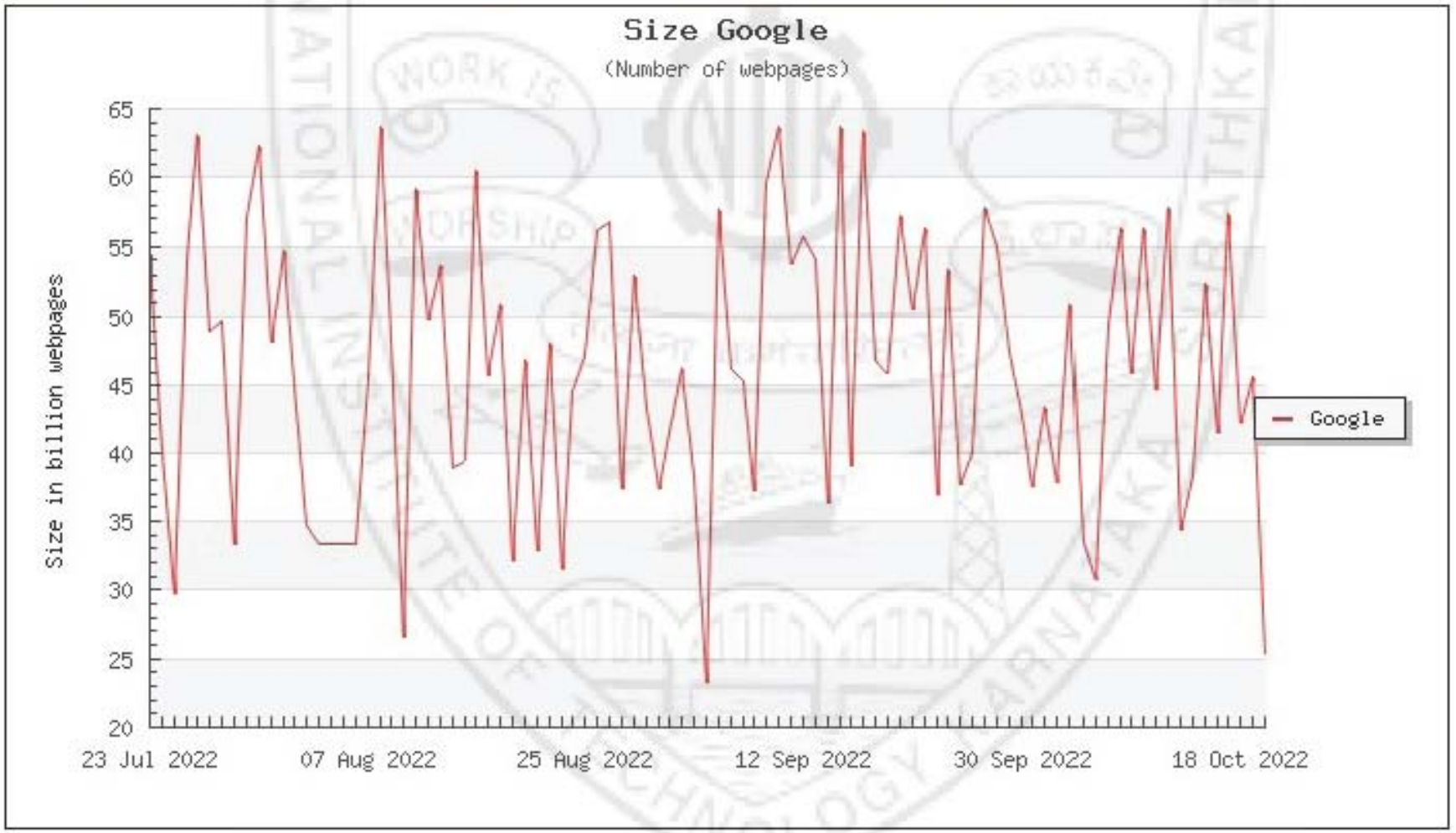
Last Month | Last Three Months | Last Year | Last Two Years | Last Five Years



Google

The size of the World Wide Web: Estimated size of Google's index

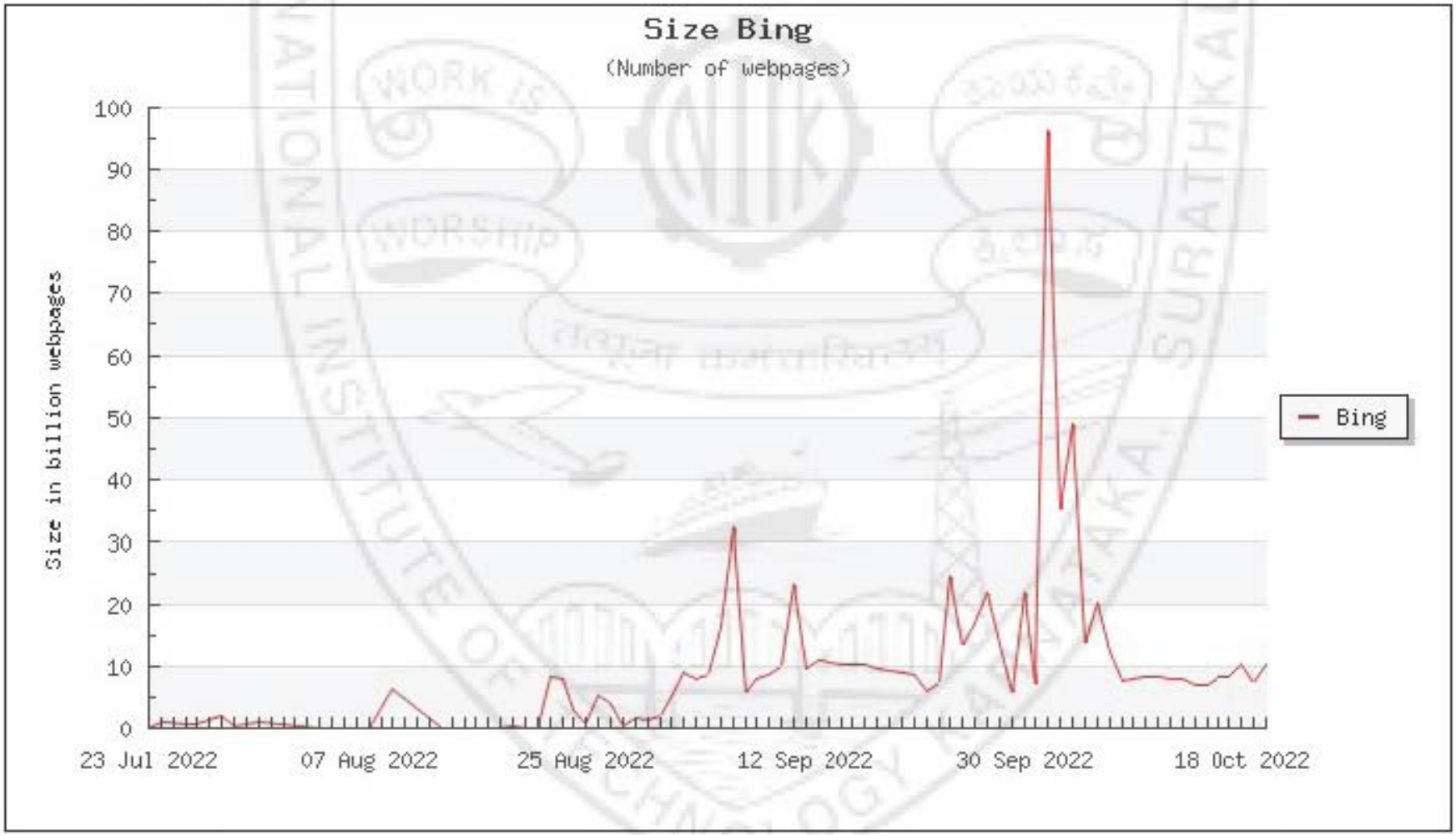
Last Month | Last Three Months | Last Year | Last Two Years | Last Five Years



bing

The size of the World Wide Web: Estimated size of Bing index

Last Month | Last Three Months | Last Year | Last Two Years | Last Five Years



Web Search: Insights

- Web contains many sources of information
  - Who to "trust"?

***Insight*: Trustworthy pages may point to each other!**

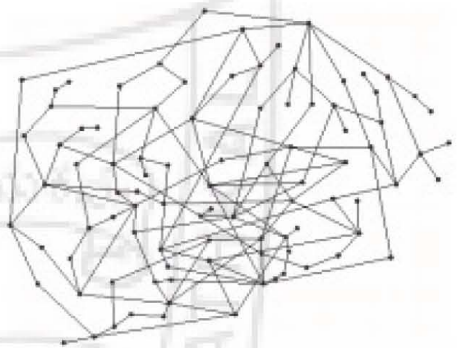
Web Search: Insights

- What is the "best" answer to queries? (For e.g. "news")
  - No single right answer

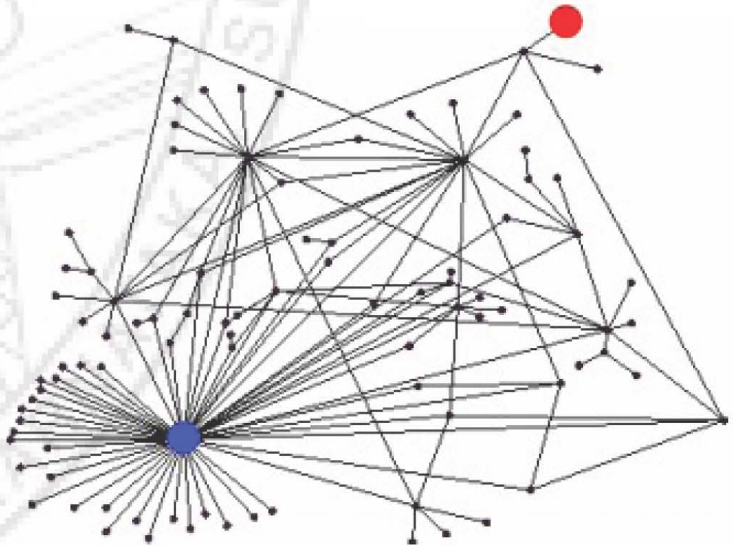
***Insight*: Pages that actually know about "news" might all be pointing to many "newspapers"**

Web Search: Insights

- All web pages are not equally "important"
  - E.g. www.mywebsite.com/homepage vs. www.stanford.edu

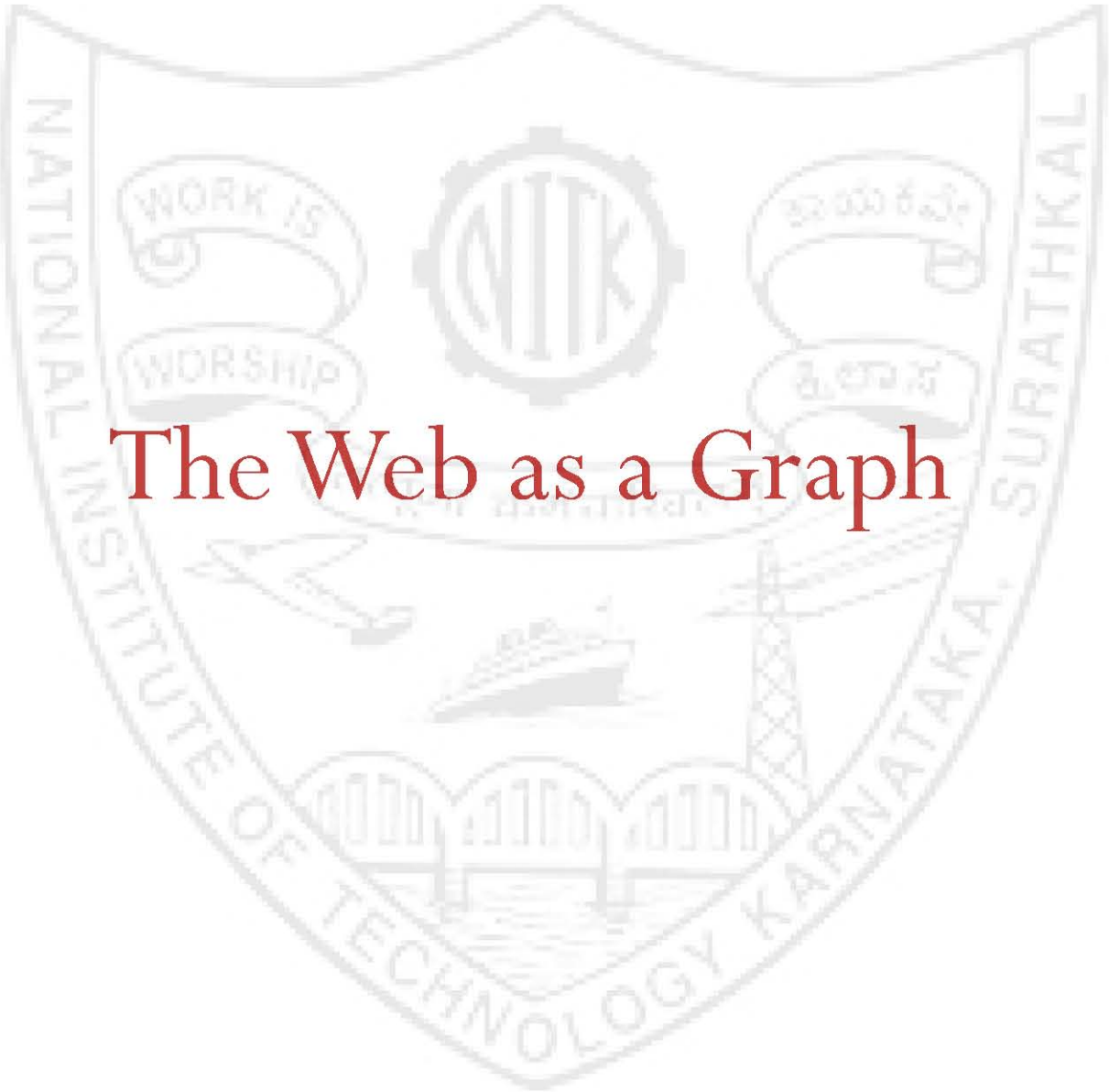


VS.



- *Fact*: There is large diversity in the web-graph node connectivity and links.

Rank the pages using web graph's ink and node structure.



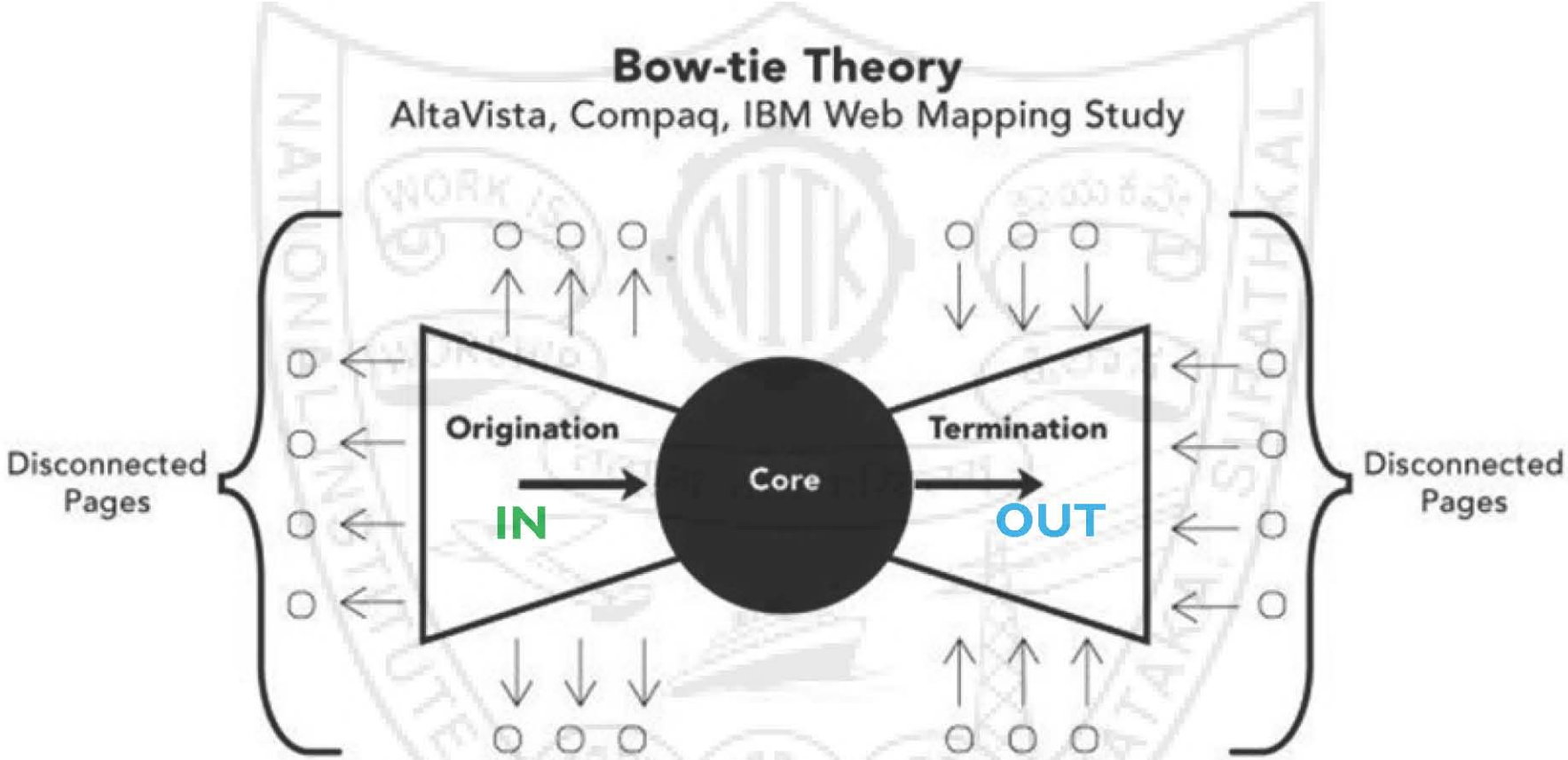
The Web as a Graph

Graph Structure of the Web

- Early large-scale study (Altavista crawls) revealed interesting properties of the Web
  - Study of 200 million nodes & 1.5 billion links
  - three sets of experiments on Web crawls between May to October 1999

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., ... & Wiener, J. (2000). Graph structure in the web. *Computer networks*, *33*(1-6), 309-320.

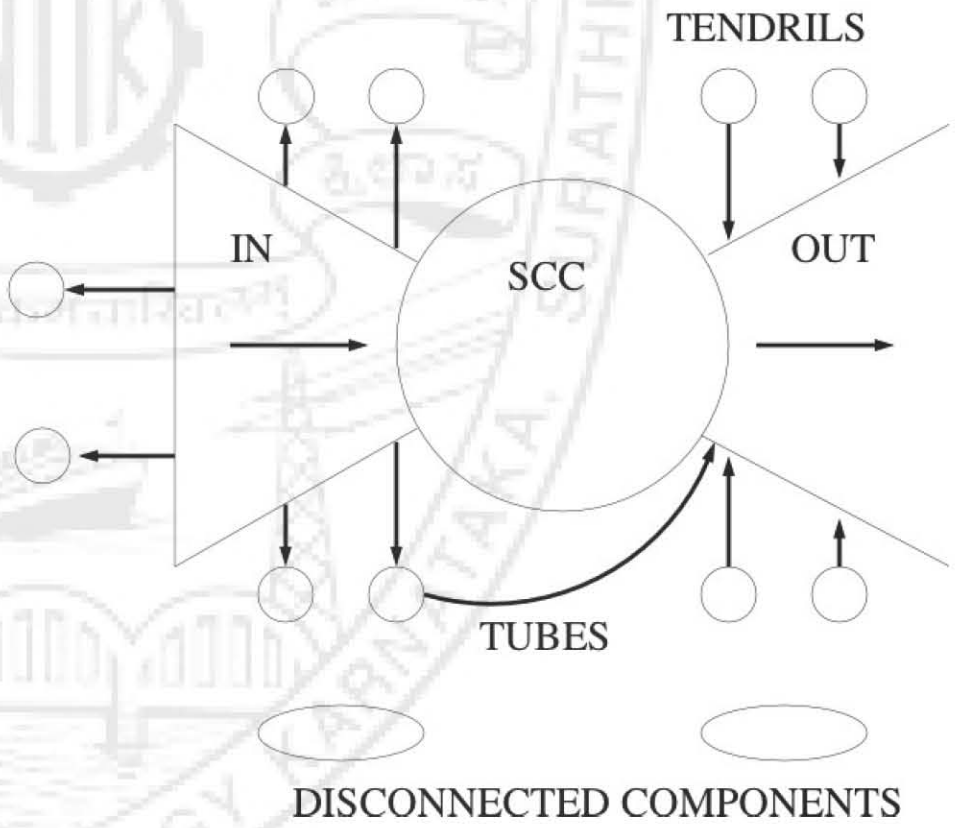
Graph Structure of the Web



Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., ... & Wiener, J. (2000). Graph structure in the web. *Computer networks*, *33*(1-6), 309-320.

Bow-tie Structure of the Web

- Strongly Connected Component (SCC)
  - Core with small-world property
- Upstream (IN)
  - Core can't reach IN
- Downstream (OUT)
  - OUT can't reach core
- Tendrils and tubes
- Disconnected

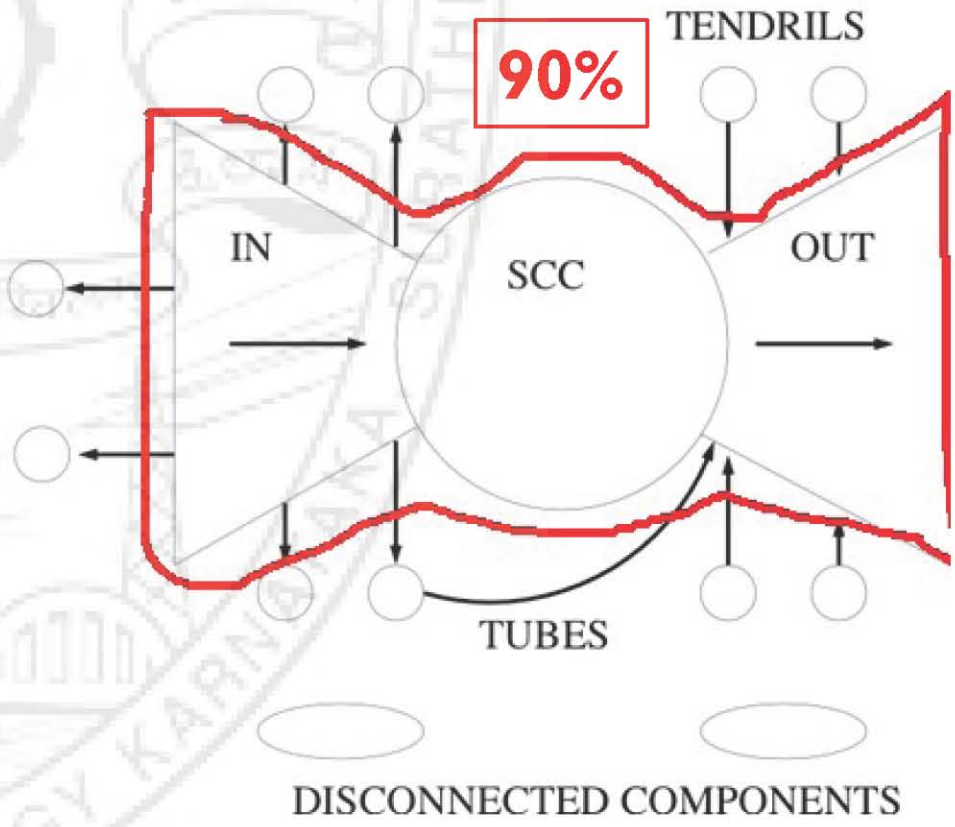


Bow-tie Structure of the Web

- Strongly Connected Component (SCC) **~47%**
  - Core with small-world property
- Upstream (IN)
  - Core can't reach IN
- Downstream (OUT)
  - OUT can't reach core

21.5%

- Tendrils and tubes **21.5%**
- Disconnected **8%**



Web Graph analysis

- Two perspectives –
  - Node-centric analysis
  - Link-centric analysis

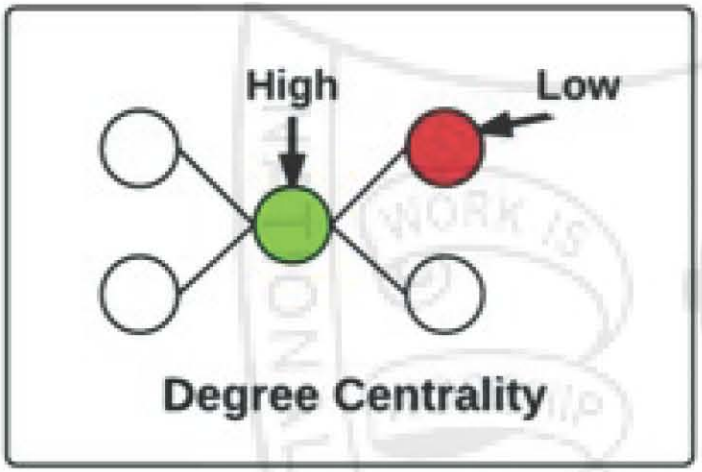
Node-centric Analysis

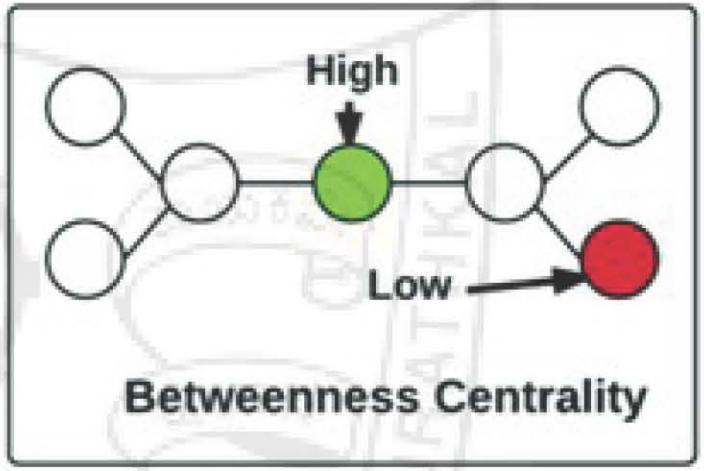
- Considering a node u in a graph, several centrality measures can be defined -
  - **Degree centrality**
  - **Betweenness centrality**
  - **Closeness centrality**
  - **Eigenvector centrality**
  - ...

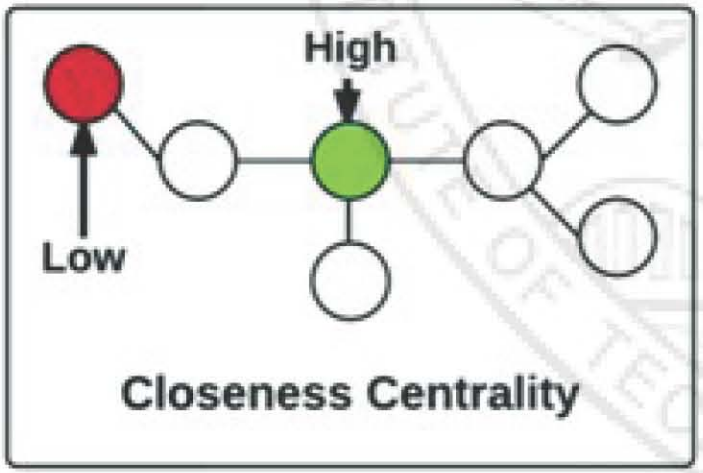
Node-centric Analysis

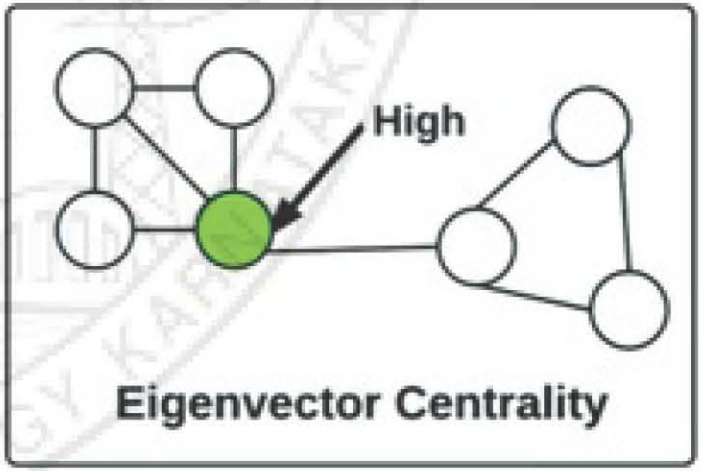
- Considering a node u in a graph, several centrality measures can be defined –
  - **Degree centrality** = degree of u
  - **Betweenness centrality** = Number of shortest paths passing through u
  - **Closeness centrality** = avg. length of shortest paths from u to all other nodes of the network.
  - **Eigenvector centrality** = measure of the influence of u in the network.
  - …

Node-centric Analysis







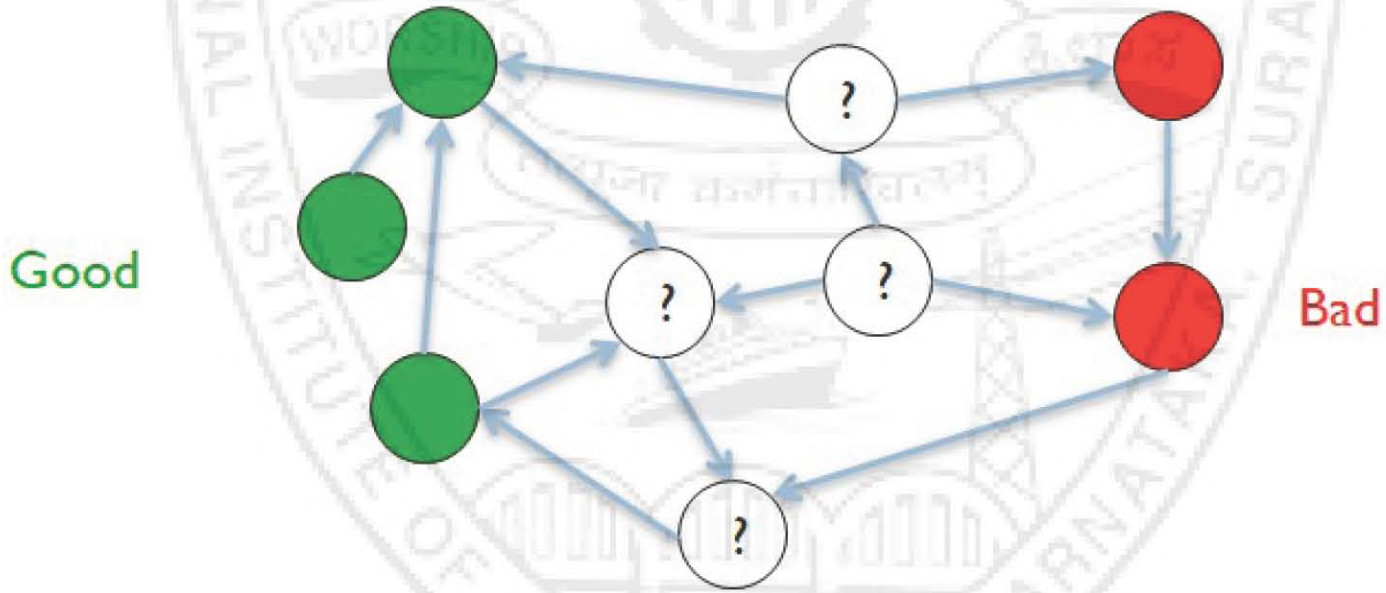


Link-centric analysis

- Links are everywhere
  - Powerful sources of authenticity and authority
    - Mail spam – which email accounts are spammers?
    - Host quality – which hosts are "bad"?
    - Phone call logs
    - …

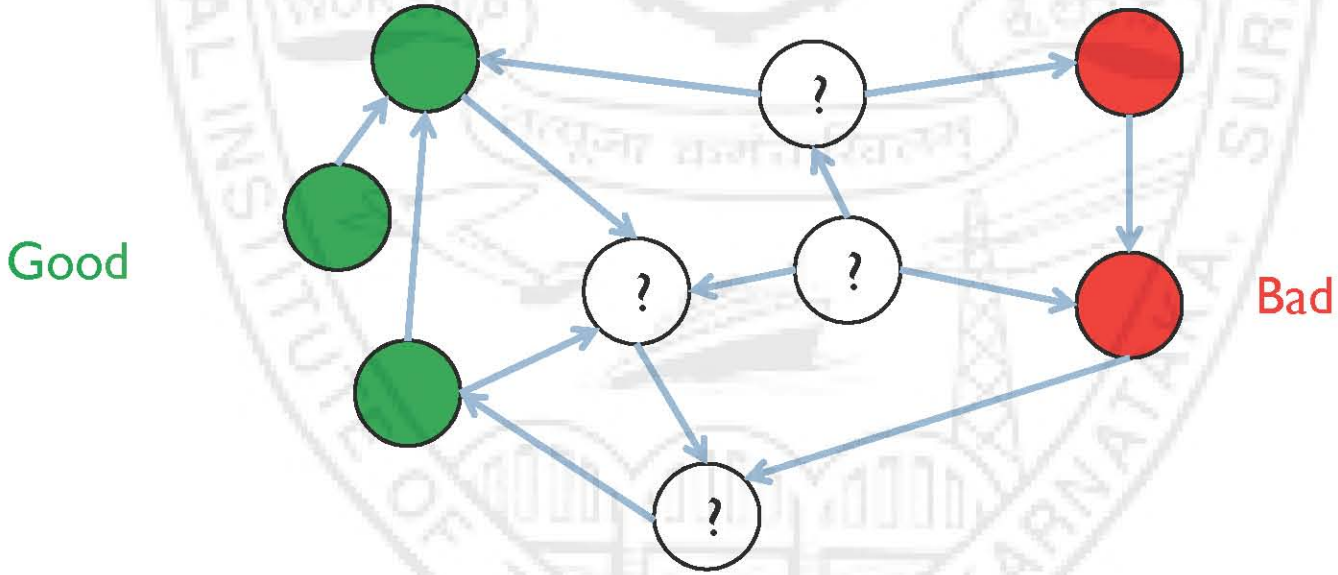
Example 1: Good/Bad/Unknown

- The Good, The Bad and The Unknown



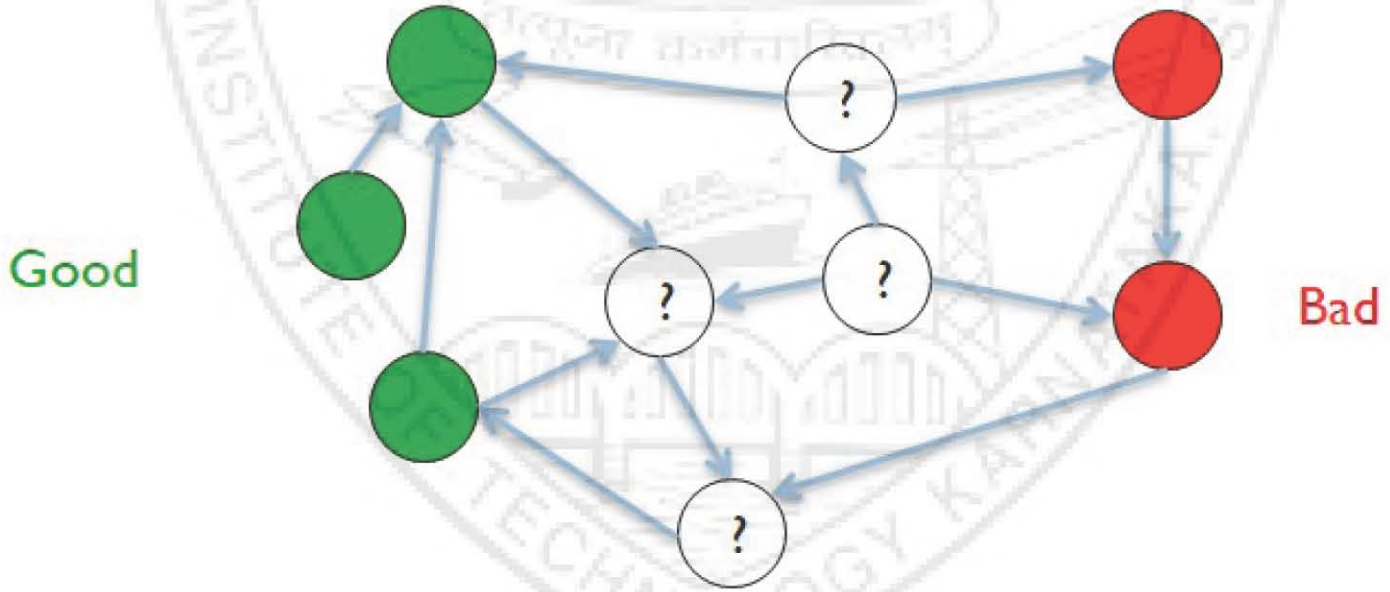
Example 1: Good/Bad/Unknown

- The Good, The Bad and The Unknown



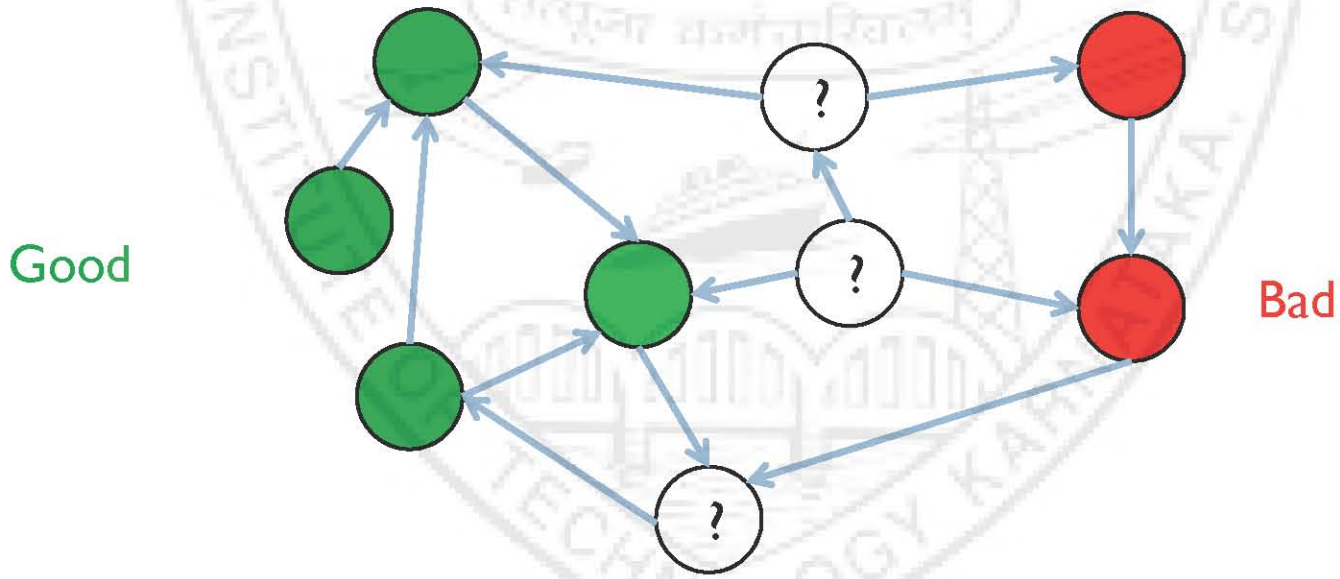
Simple iterative logic

- Good nodes won't point to Bad nodes
  - If you point to a Bad node, you're Bad
  - If a Good node points to you, you're Good



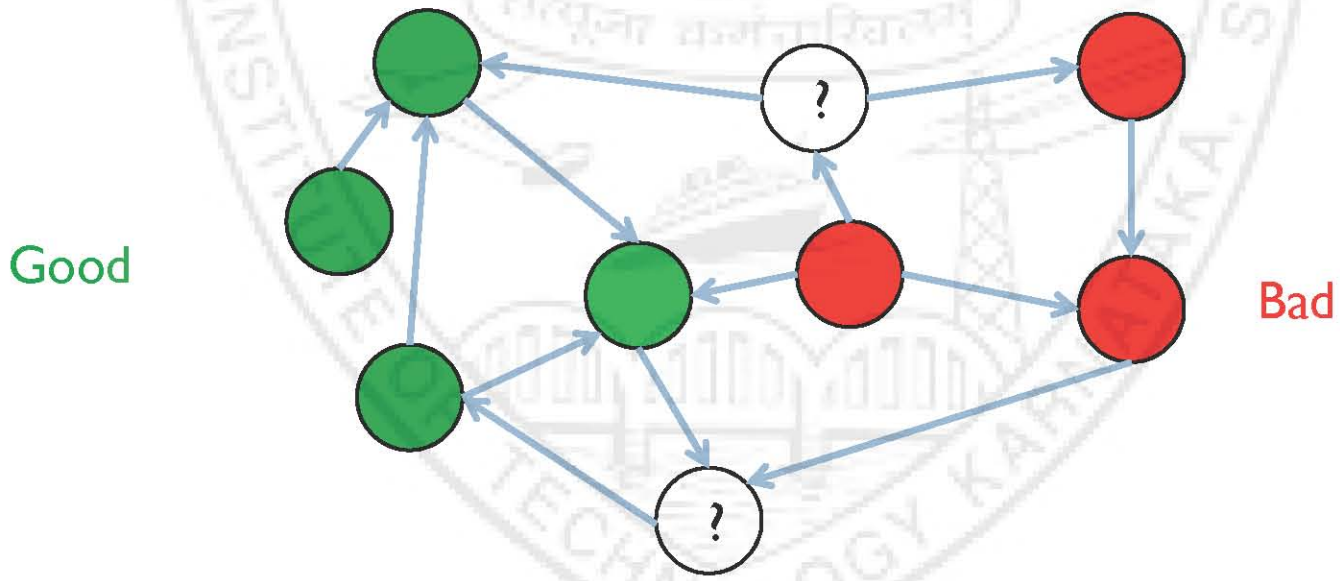
Simple iterative logic

- Good nodes won't point to Bad nodes
  - If you point to a Bad node, you're Bad
  - If a Good node points to you, you're Good



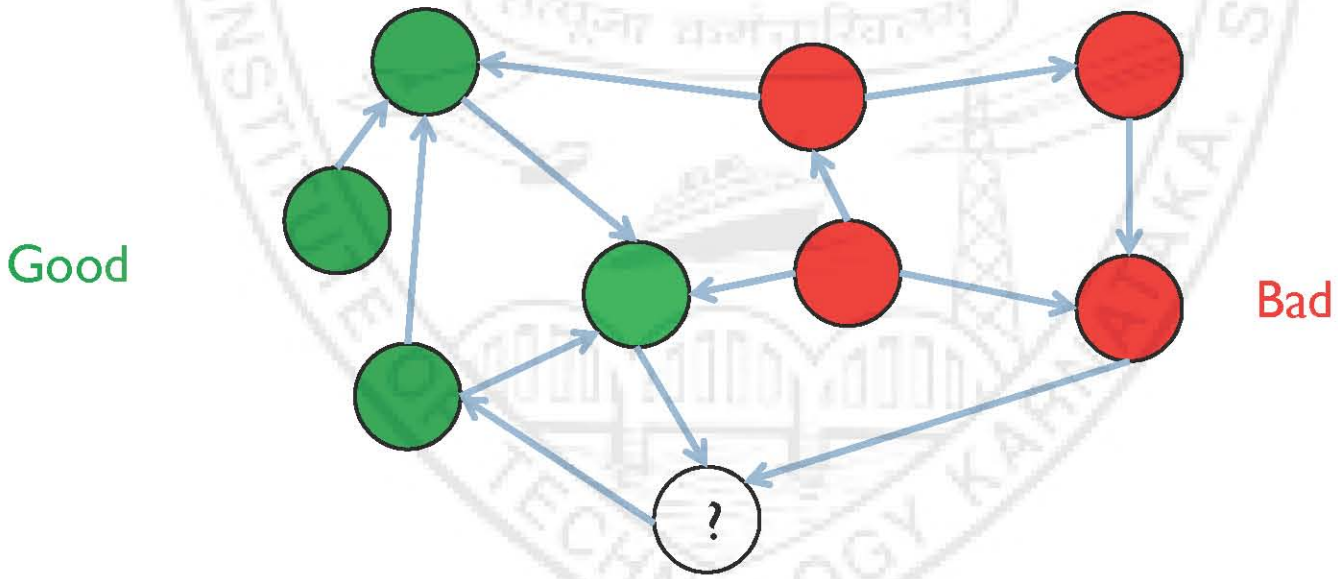
Simple iterative logic

- Good nodes won't point to Bad nodes
  - If you point to a Bad node, you're Bad
  - If a Good node points to you, you're Good



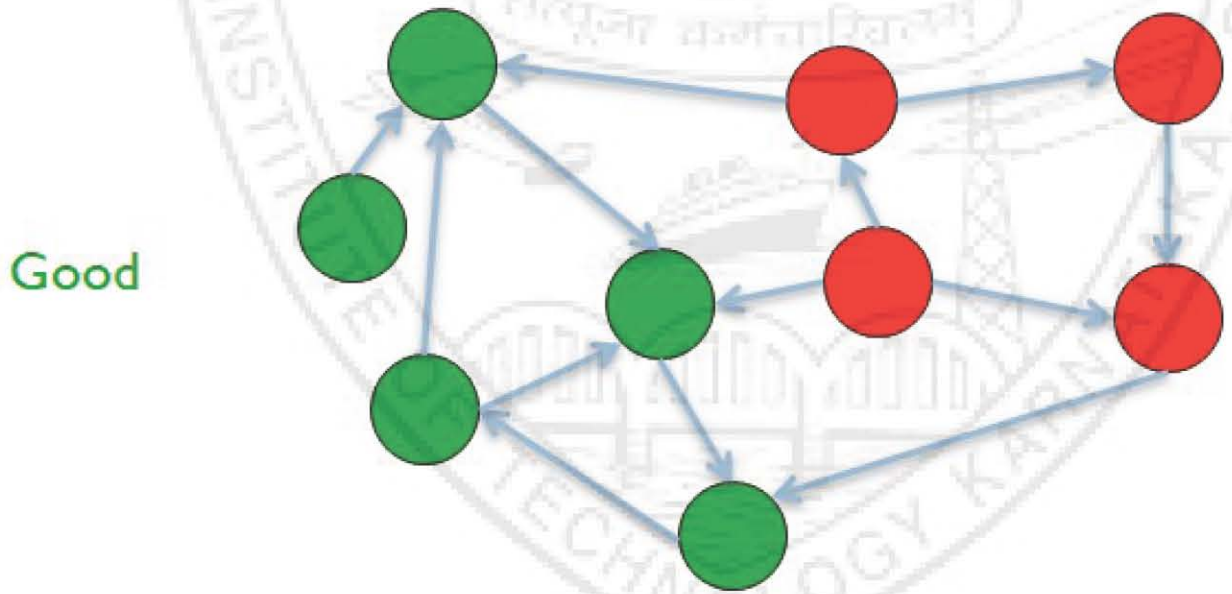
Simple iterative logic

- Good nodes won't point to Bad nodes
  - If you point to a Bad node, you're Bad
  - If a Good node points to you, you're Good



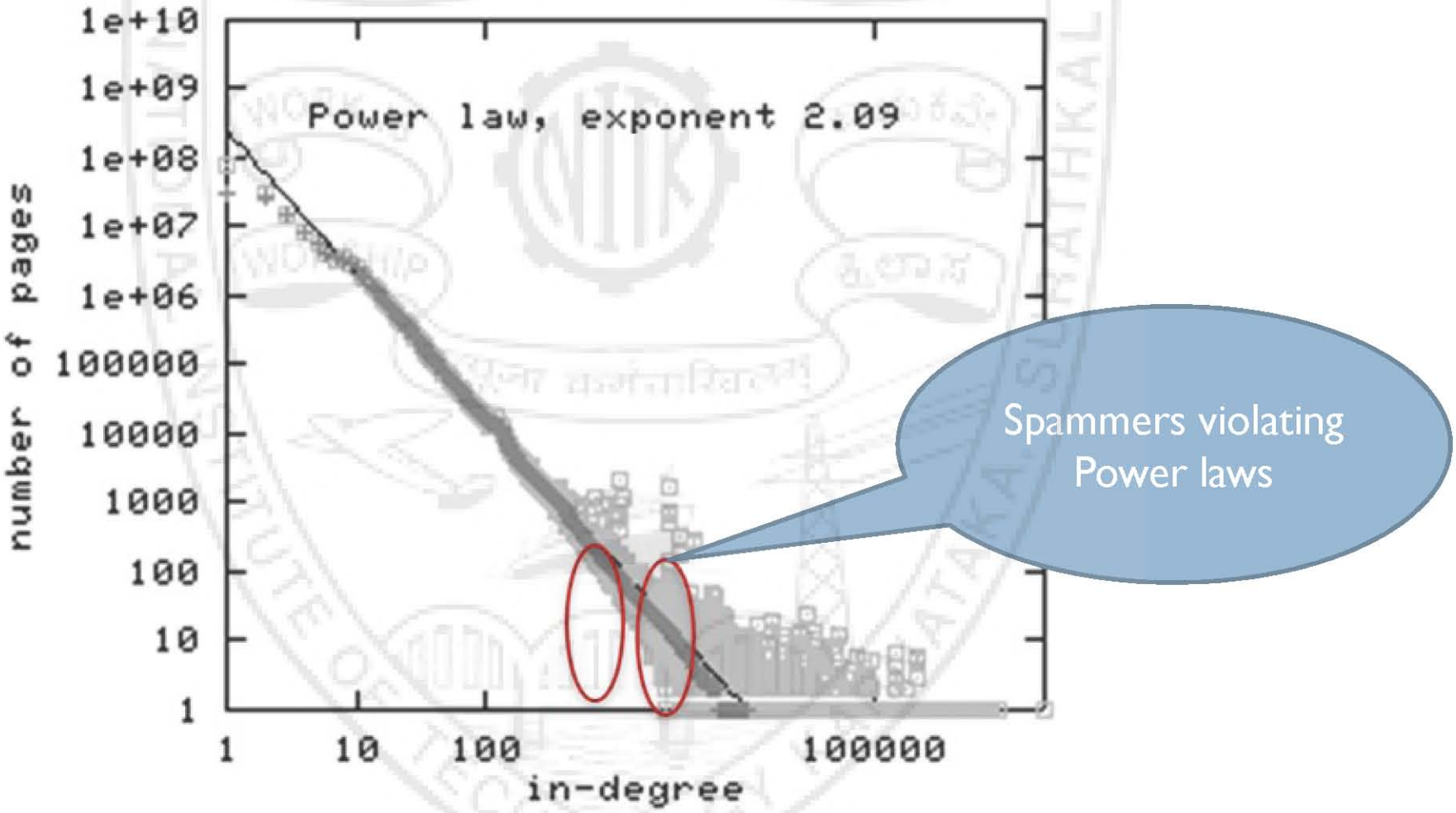
Simple iterative logic

- Good nodes won't point to Bad nodes
  - If you point to a Bad node, you're Bad
  - If a Good node points to you, you're Good



Sometimes need probabilistic analysis – e.g., mail spam

Example 2: In-links to pages – unusual patterns



Importance of Link Analysis

- Link analysis is fundamental to most IR system functionalities –
  - Scoring and ranking
  - Link-based clustering – topical structure from links
  - Links as features in classification – documents that link to one another are likely to be on the same subject
  - Crawling - Based on the links seen, where do we go next?

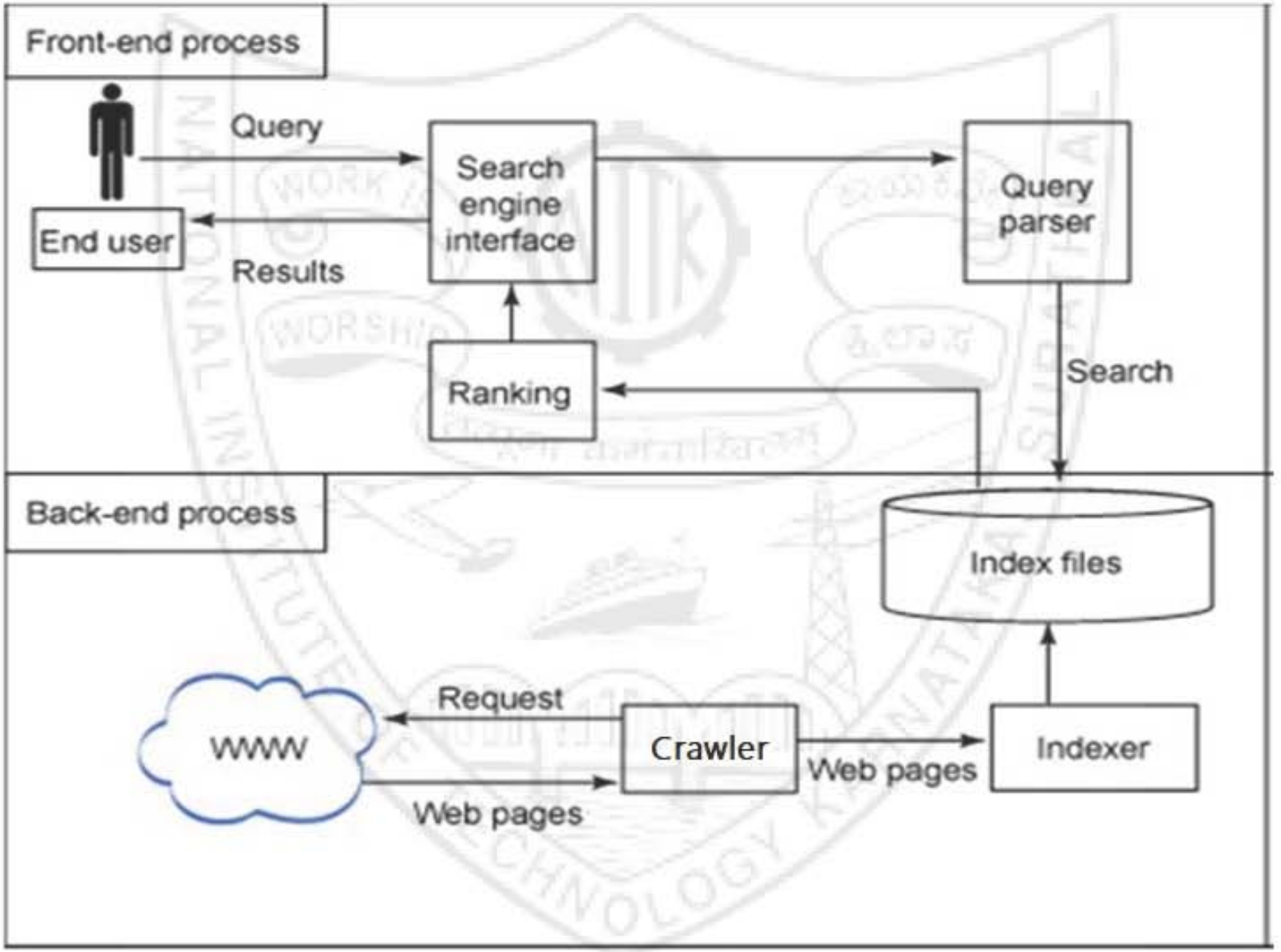
Web Crawlers

Working of a Search Engine

- **Terms used -** Spider, Crawler, Indexer, Bot, Search Algorithm.

- **Common Characteristics:**
  - Find matching documents and display them according to relevance.
  - Frequent updates to proprietary ranking algorithm.
  - Strive to produce "better", more relevant results than competitors.

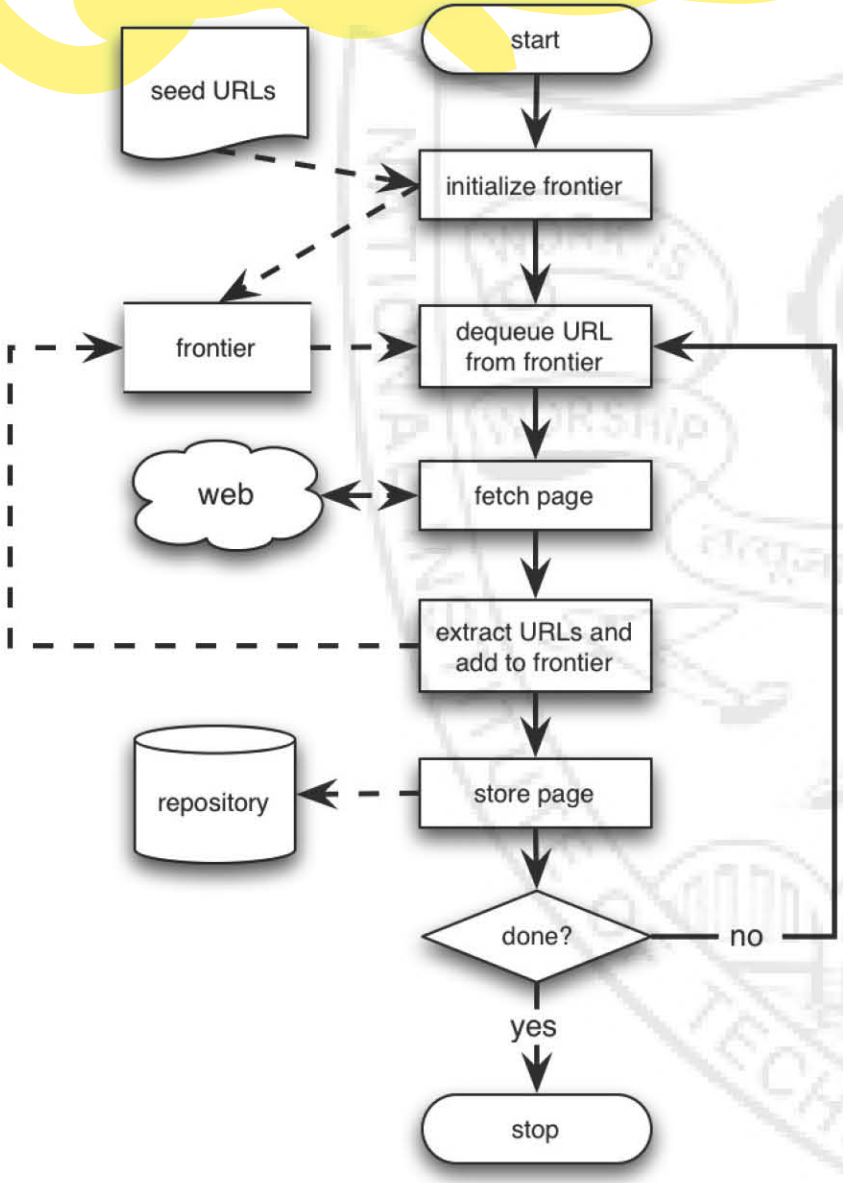
Working of a Search Engine



Working of a Search Engine

- Spider "**crawls**" the web to find new documents (web pages, other documents) typically by following hyperlinks from websites already in their database.
- Search engine "**indexes**" the content in these documents by adding it to its databases and then **periodically updates** this content.
- Search engines "**search their own indexed databases**" **when a user enters in a search** to find related documents (not searching web pages in real-time).
- Search engines "**rank**" **the resulting documents** using an algorithm *(mathematical formula)* by assigning various weights and **ranking factors**

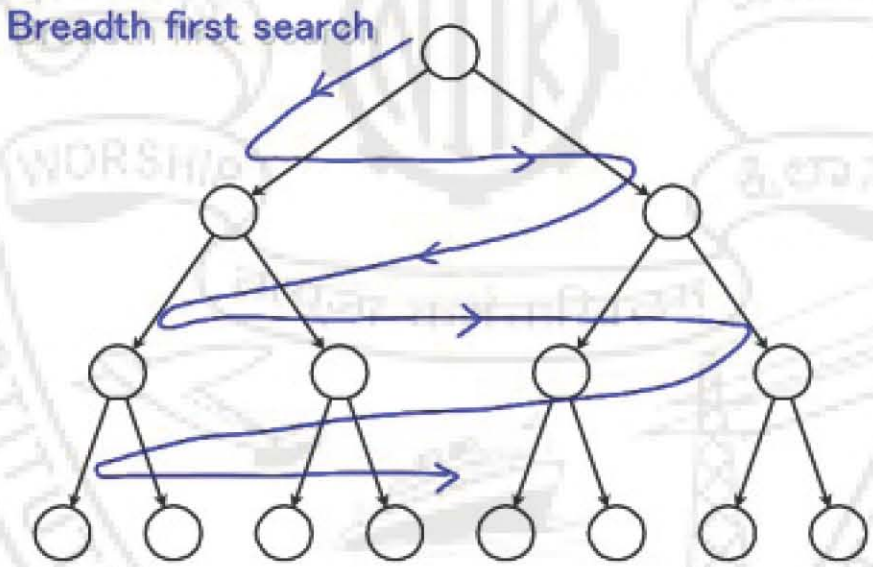
Web Crawler -- *working*



- a sequential crawler
- Seeds can be any list of starting URLs
- Order of page visits is determined by frontier data structure
- Stop criterion can be anything

Crawlers - Web Graph travesal

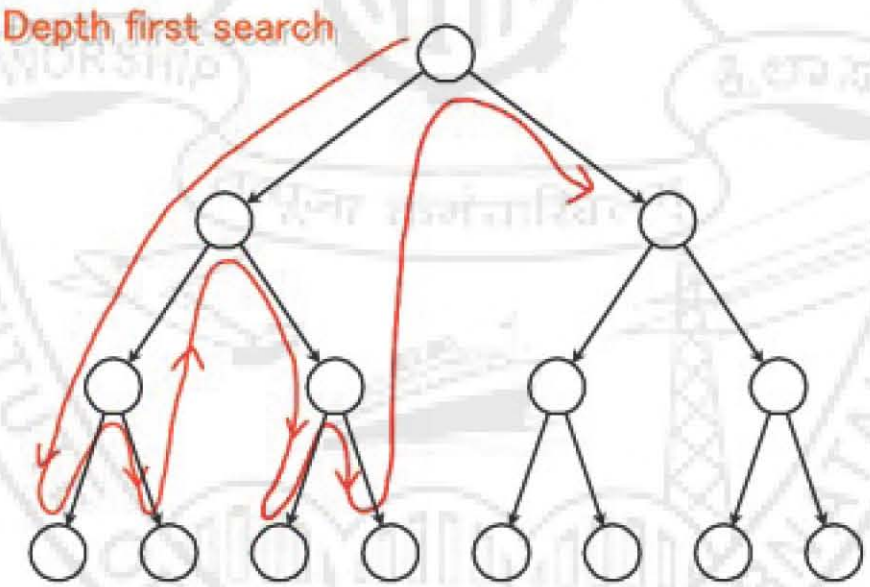
- Breadth First Search
  - Implemented with QUEUE-type mechanisms (FIFO)



  - Finds pages along shortest paths
    - If we start with "good" pages, this keeps us close; maybe also to other good stuff…

Crawlers - Web Graph travesal

- Depth First Search
  - Implemented with STACK type mechanisms (LIFO)

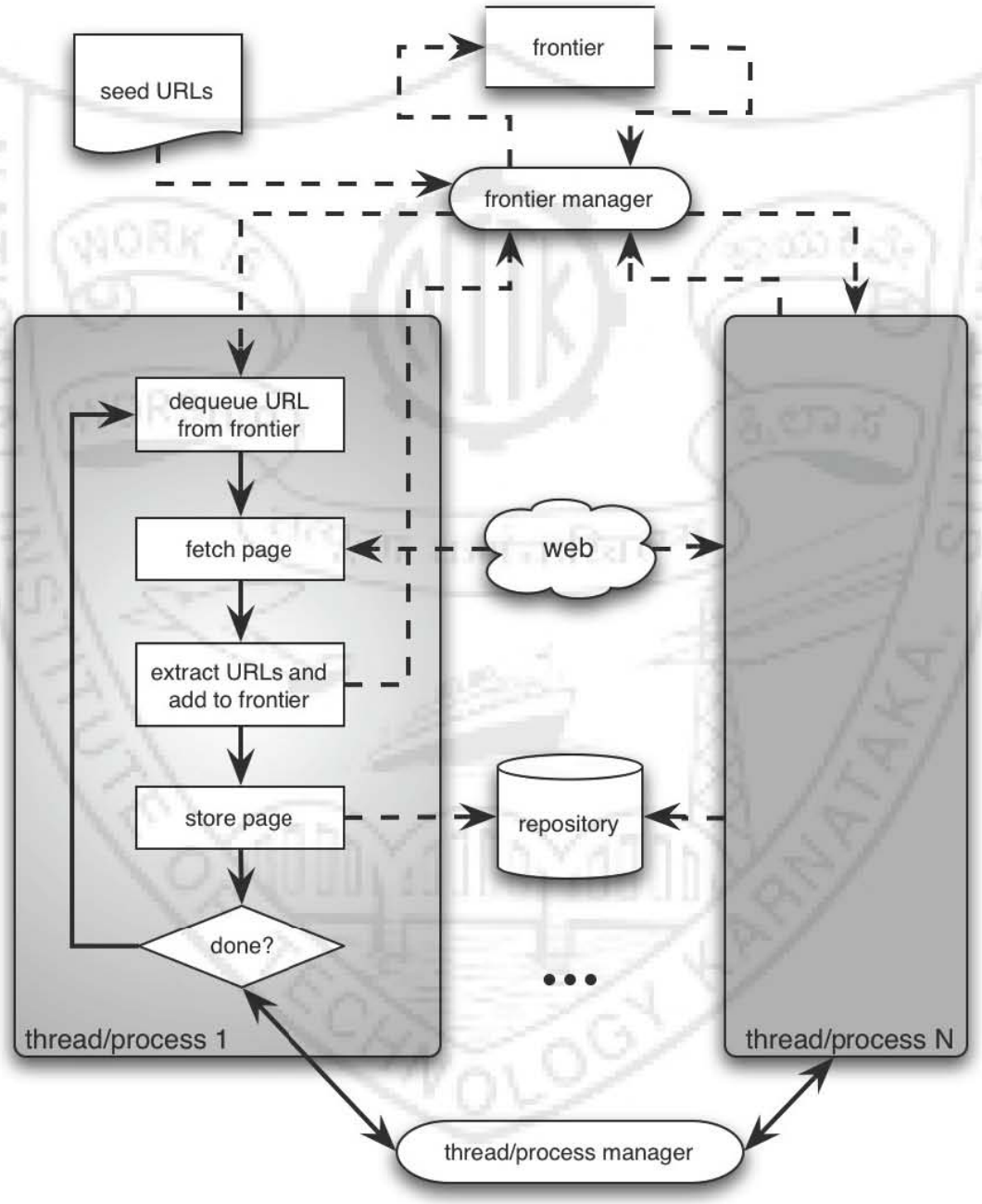


  - Can cause one to wander away ("lost in cyberspace")

Concurrent crawlers

- Can use multi-processing or multi-threading
  - Each process or thread works like a sequential crawler, except they share data structures: *frontier* and *repository*
    - Shared data structures must be synchronized *(locked for concurrent writes)*
  - Achieve a speedup factor of 5-10 !!

Architecture of a concurrent crawler



Preferential crawlers

- Assuming each page has an importance measure, *I(p)*, if -
  - Want to visit pages in order of decreasing *I(p)*
  - Maintain the frontier as a priority queue sorted by *I(p)*

- *Requirement:* Selective bias toward some pages,
  - E.g. most "relevant"/topical, closest to seeds, most popular/largest PageRank, unknown servers, highest rate/amount of change, etc…

- *Solution*: preferential crawlers

Preferential crawlers – Types

- Focused crawlers
  - Supervised learning -- classifier based on labeled examples
- Topical crawlers
  - Best-first search based on similarity (topic, parent)
- Adaptive crawlers
  - Reinforcement learning
  - Evolutionary algorithms

More reading

- Najork, Marc. "Web Crawler Architecture." (2009): 3462-3465.

- Shkapenyuk, V., & Suel, T. (2002, February). Design and implementation of a high-performance distributed web crawler. In *Proceedings 18th International Conference on Data Engineering* (pp. 357-368). IEEE.

- Chakrabarti, Soumen, Martin Van den Berg, and Byron Dom. "Focused crawling: a new approach to topic-specific Web resource discovery." *Computer networks* 31.11-16 (1999): 1623-1640.