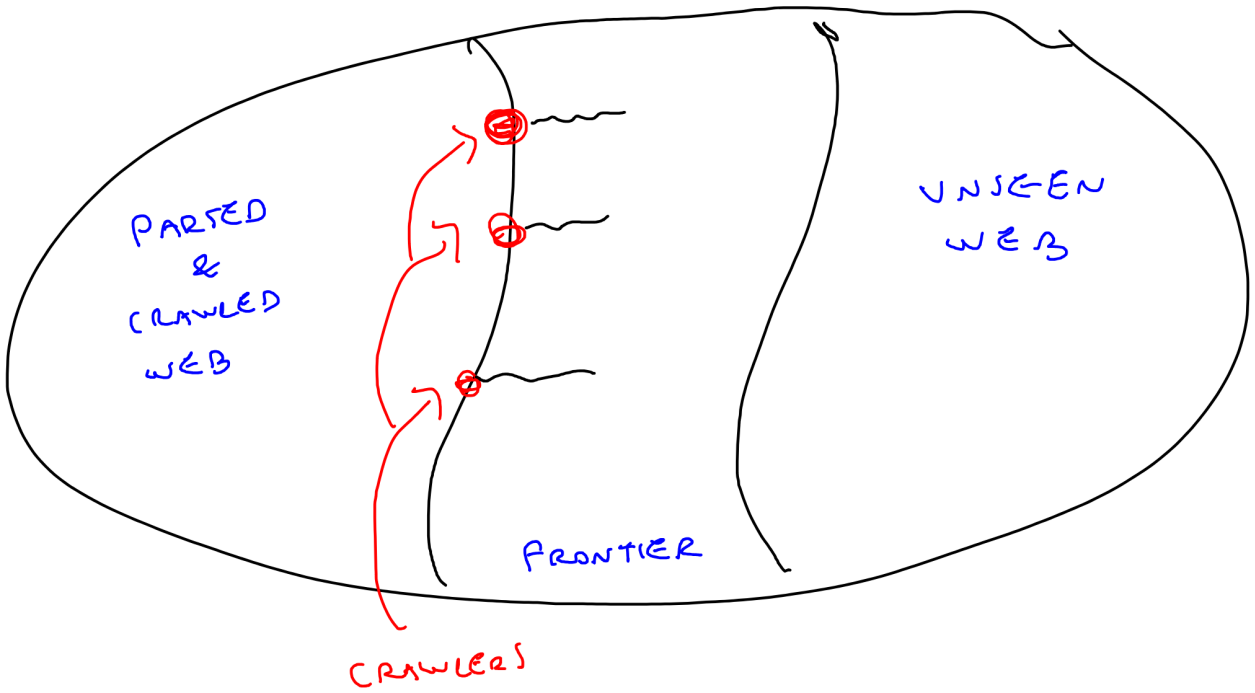
CRAWLING

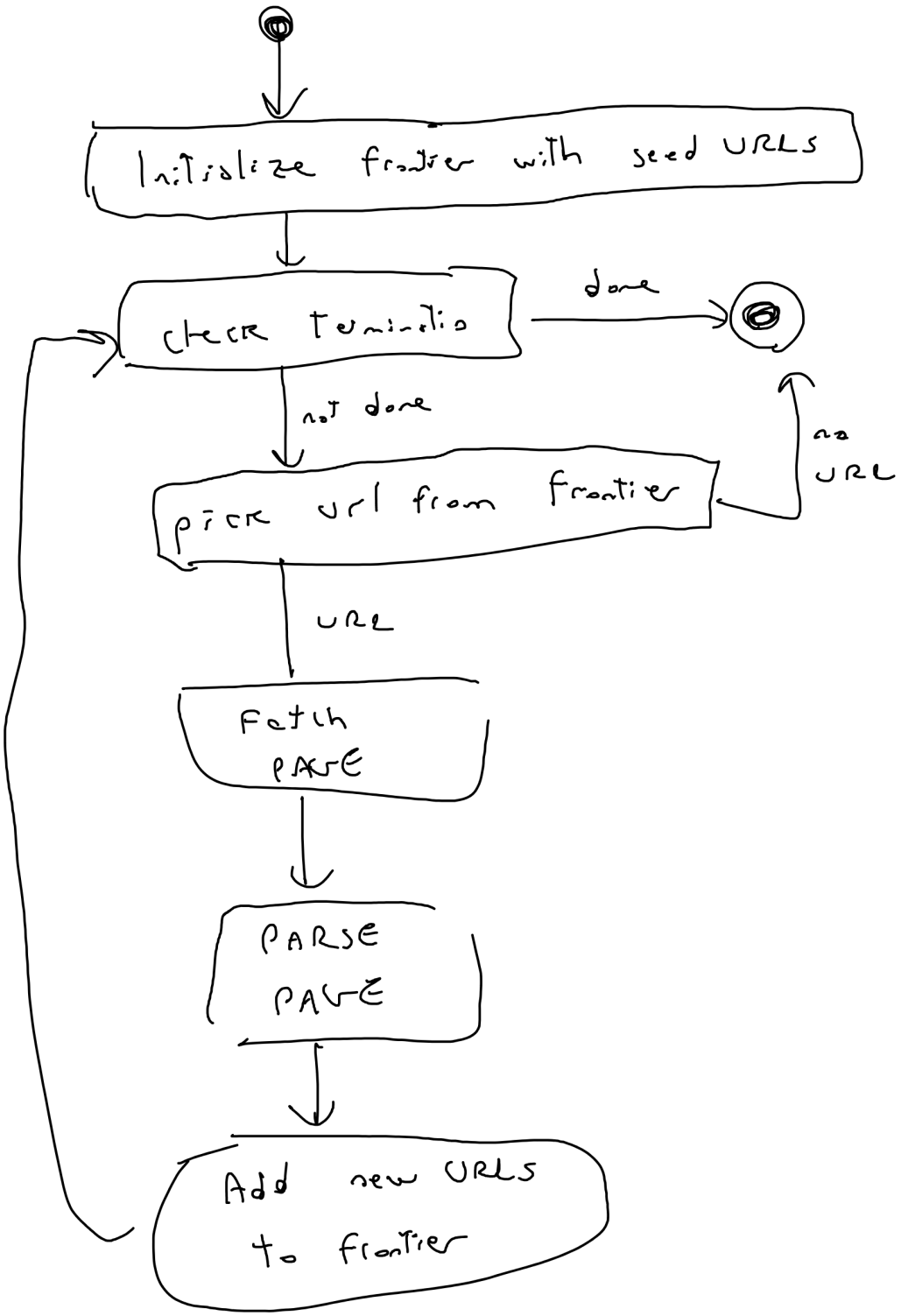
WALKING OVER THE WEB GRAPH IN ORDER TO TRACE ITS ROUTES

CHALLENGES

- You should crawl the best pages first (Quality)
- You should avoid duplication or near duplication (Efficiency)
- Minimize the load of the crawled hosts (Etiquette)
- Avoid Spam pages and spider traps (Malicious pages)
- How to cover the whole web (Coverage)
- How much our competitors are able to cover in respect with our coverge (Relative coverage)
- How often should I crawl (frequency)



CRAWLER WORKFLOW



Several metrics to pick up the best URL from the frontier : Random, BFS, DFS, Topic driven, Random, PageRank - - -

↳ how popular is the page