EPL 602: Lab 3

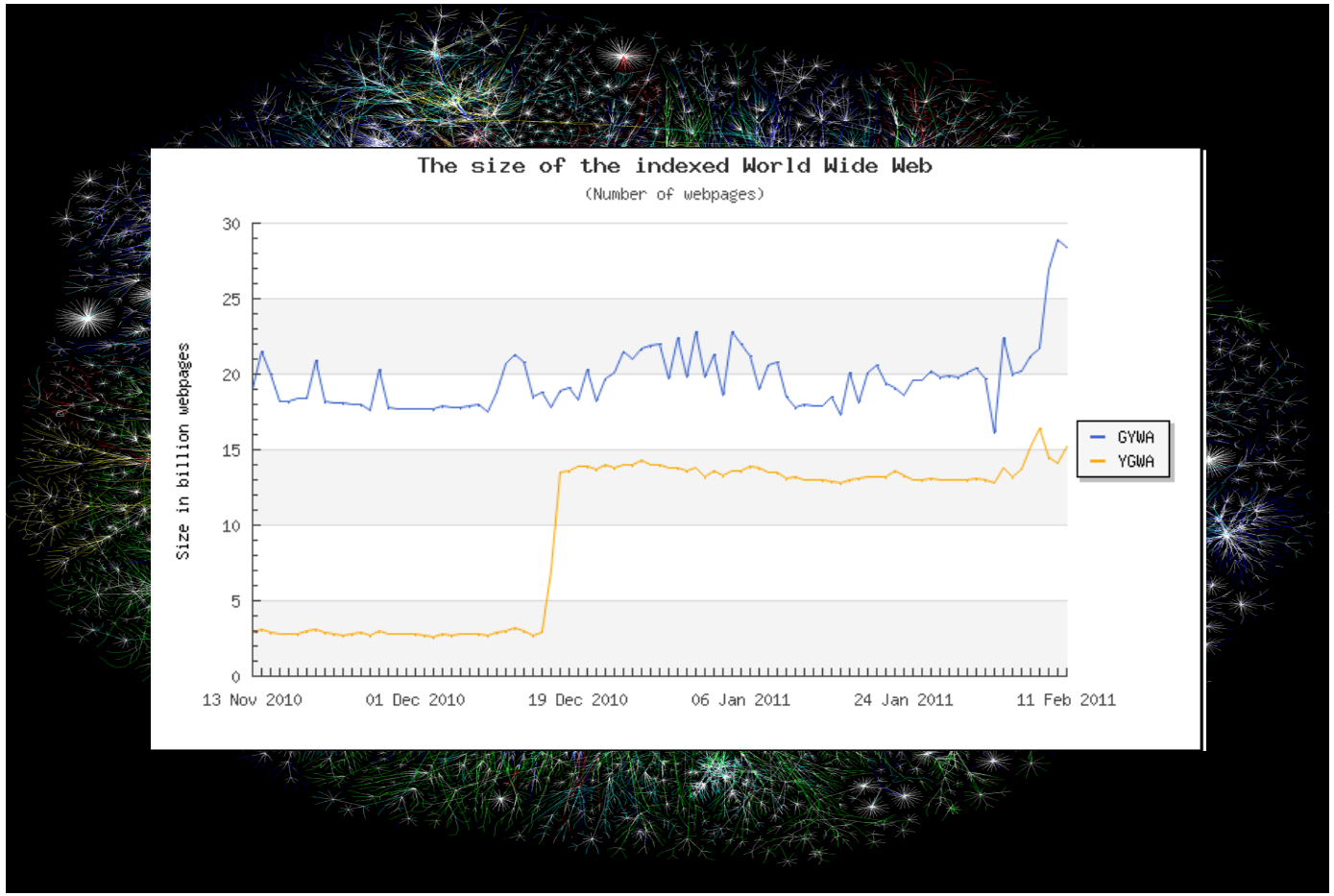
Assignment 1: A Web Crawler

University of Cyprus
Department of
Computer Science

Andreas Kamilaris

The World Wide Web Map

University of Cyprus



How can Google/Yahoo predict the size of the Web?

Web Crawlers

University of Cyprus

- A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner, in an orderly fashion.
- It is one type of bot, or software agent.

- It starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier.
- URLs from the frontier are recursively visited according to a set of policies (updates).

1st Programming Exercise

University of Cyprus

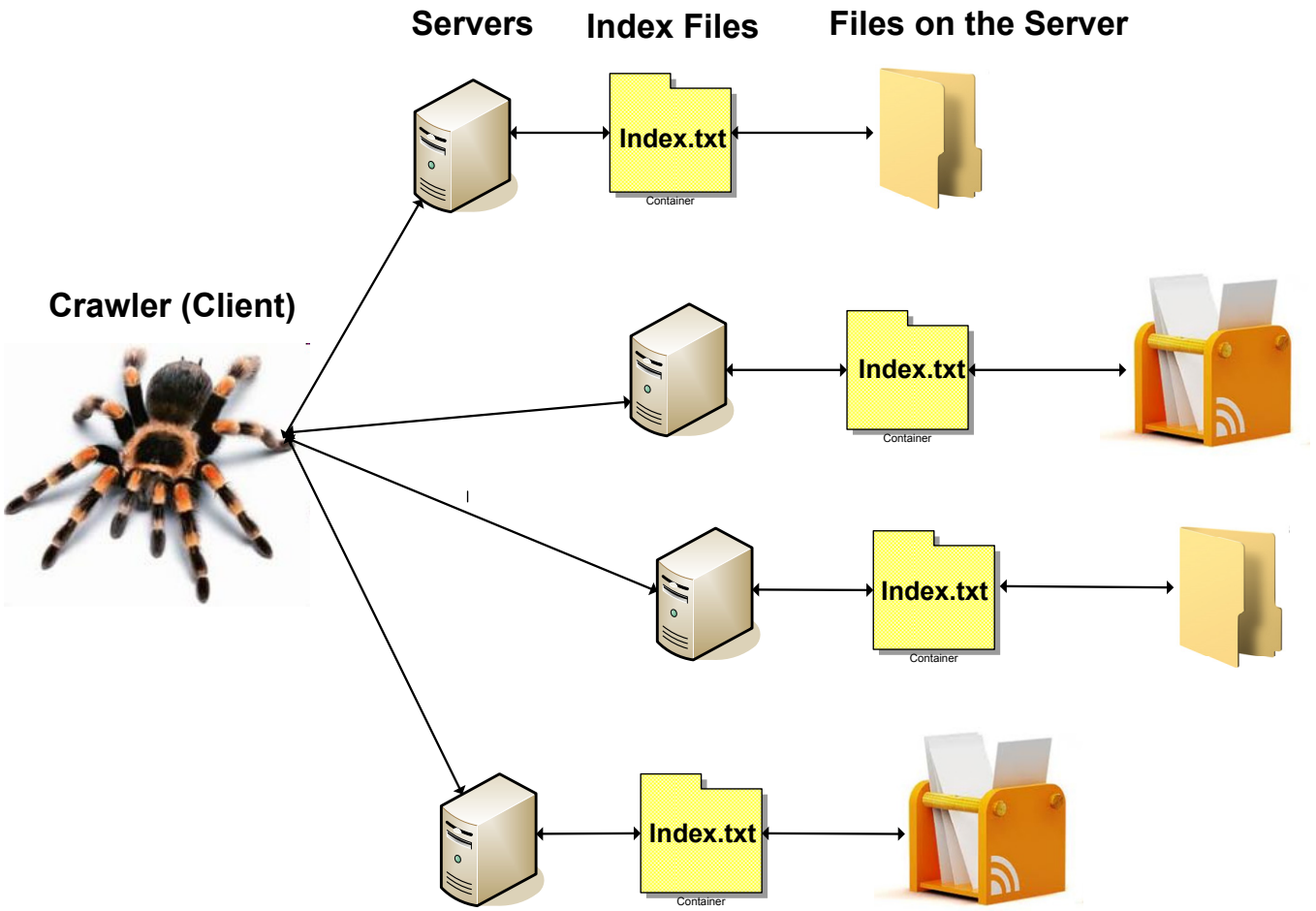
- Create a simple Web crawler!
- The main difference from real Web crawlers is that instead of websites, your crawler will discover files.
- Get experience in Socket programming and Internet primitive communication.
- Use as a programming language Java or C.

Crawler Operation

1. The Web crawler will act as a client, searching for servers which may contain files.
2. The crawler will receive the addresses of these servers from a local file.
3. When the crawler contacts a server, it must ask for the index.txt file, which would contain the filenames that are stored on that server.
4. The crawler must then contact the server consecutively to receive the contents of all files.
5. The crawler must print every file on screen.

Crawler Operation

University of Cyprus



Notes

1. You need to write code for a simple client (crawler) and a simple server, which would communicate through the Internet.
2. You can assume that the local file that contains server addresses contains only one address.
3. Local Vs Remote servers.
4. The server undestands two types of commands:
   - GET index.txt
   - GET {file}

Deliverables

University of Cyprus

- Deadline is 18th February 2011 23:59.
- You need to include:
  - Source code with comments.
  - A Brief Documentation.
- E-mail submission including a zip attachment.
- Assistant's Email: kami@cs.ucy.ac.cy