

Busca, Recuperação e Mineração na Web

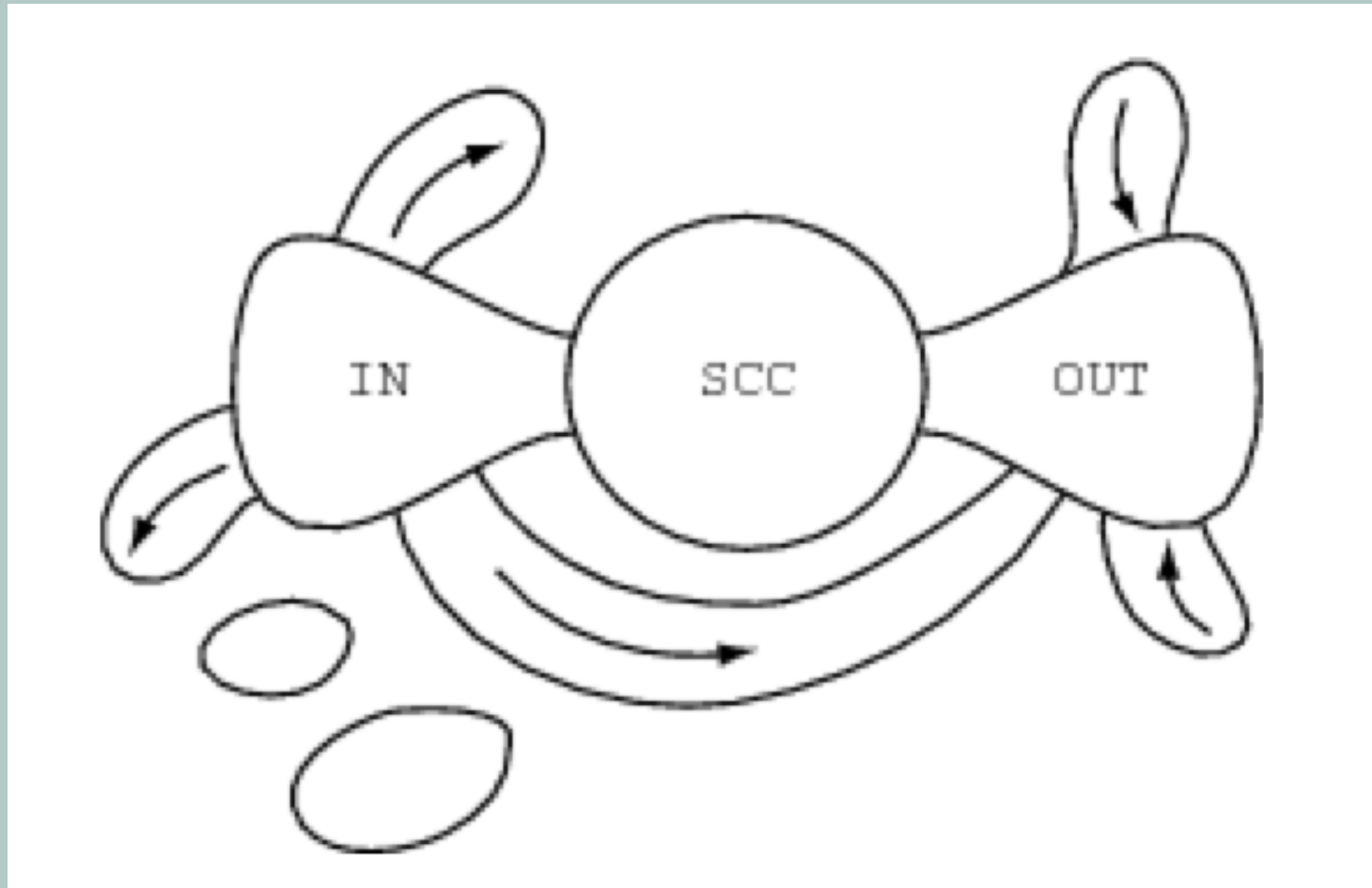
Carlos Bazilio

carlosbazilio@id.uff.br

<https://carlosbazilio.github.io>

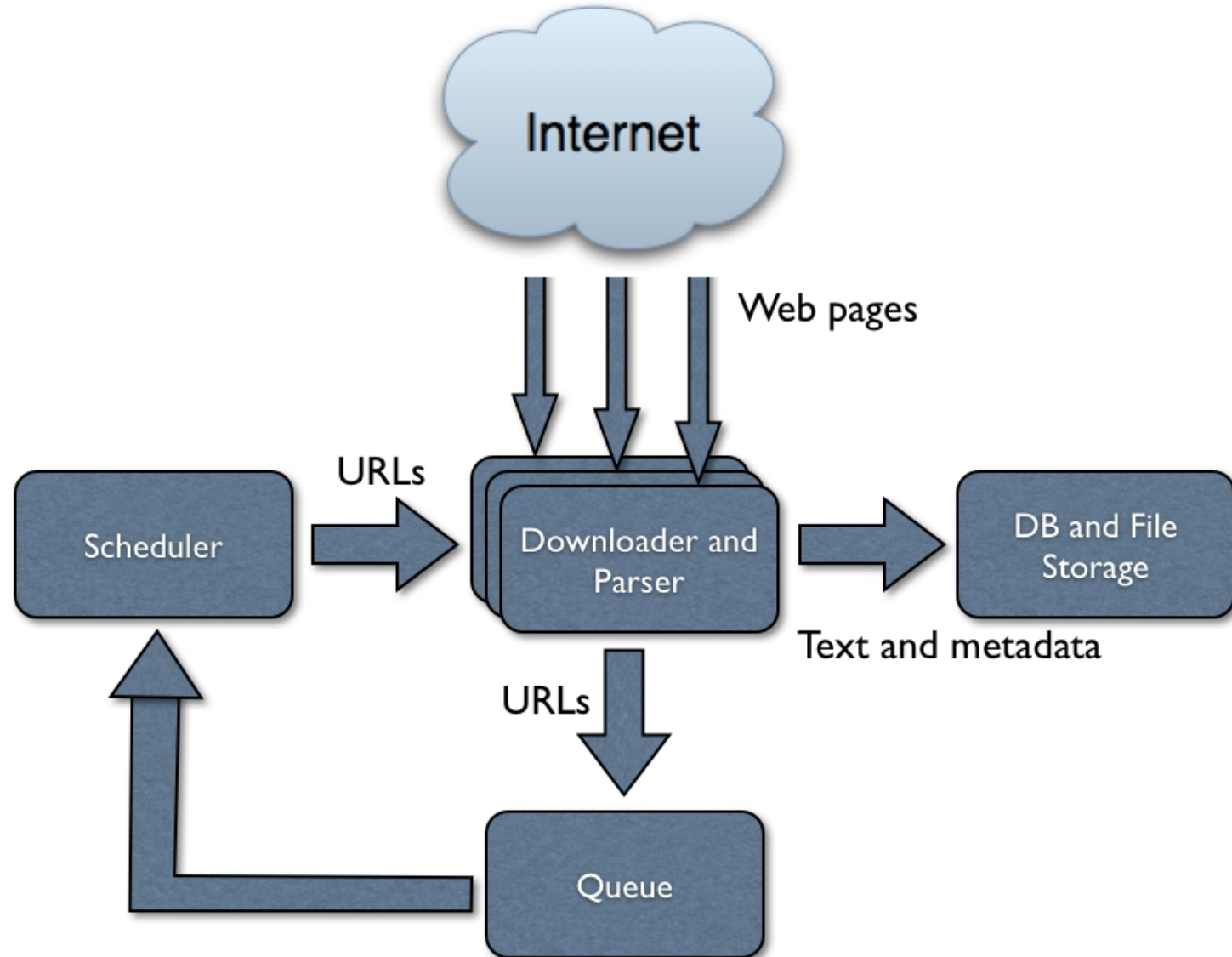
Depto de Computação
Instituto de Ciência e Tecnologia
Universidade Federal Fluminense

Estrutura do Grafo Web

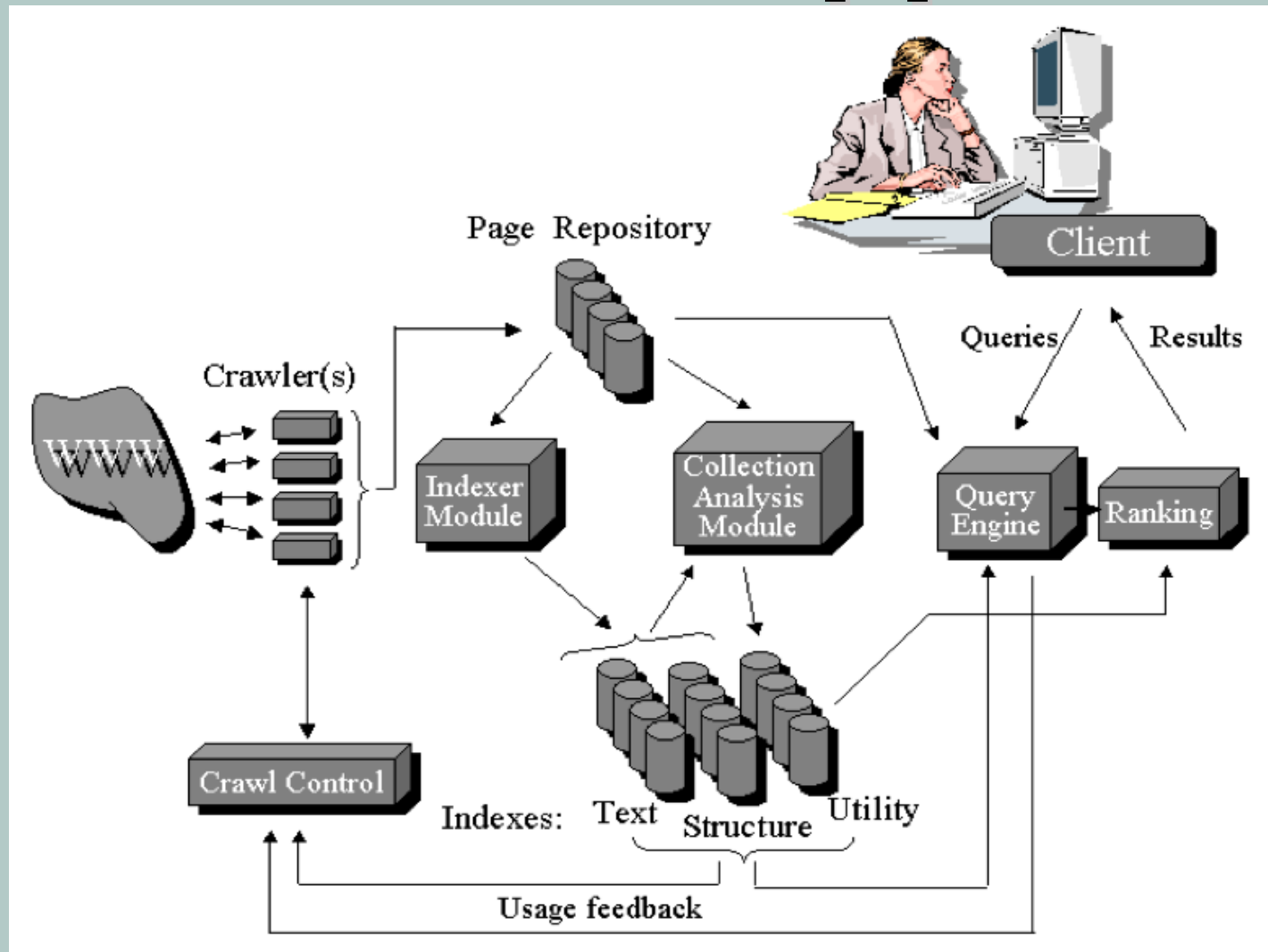


<https://www.cs.cornell.edu/home/kleinber/networks-book/> (Cap. 13)

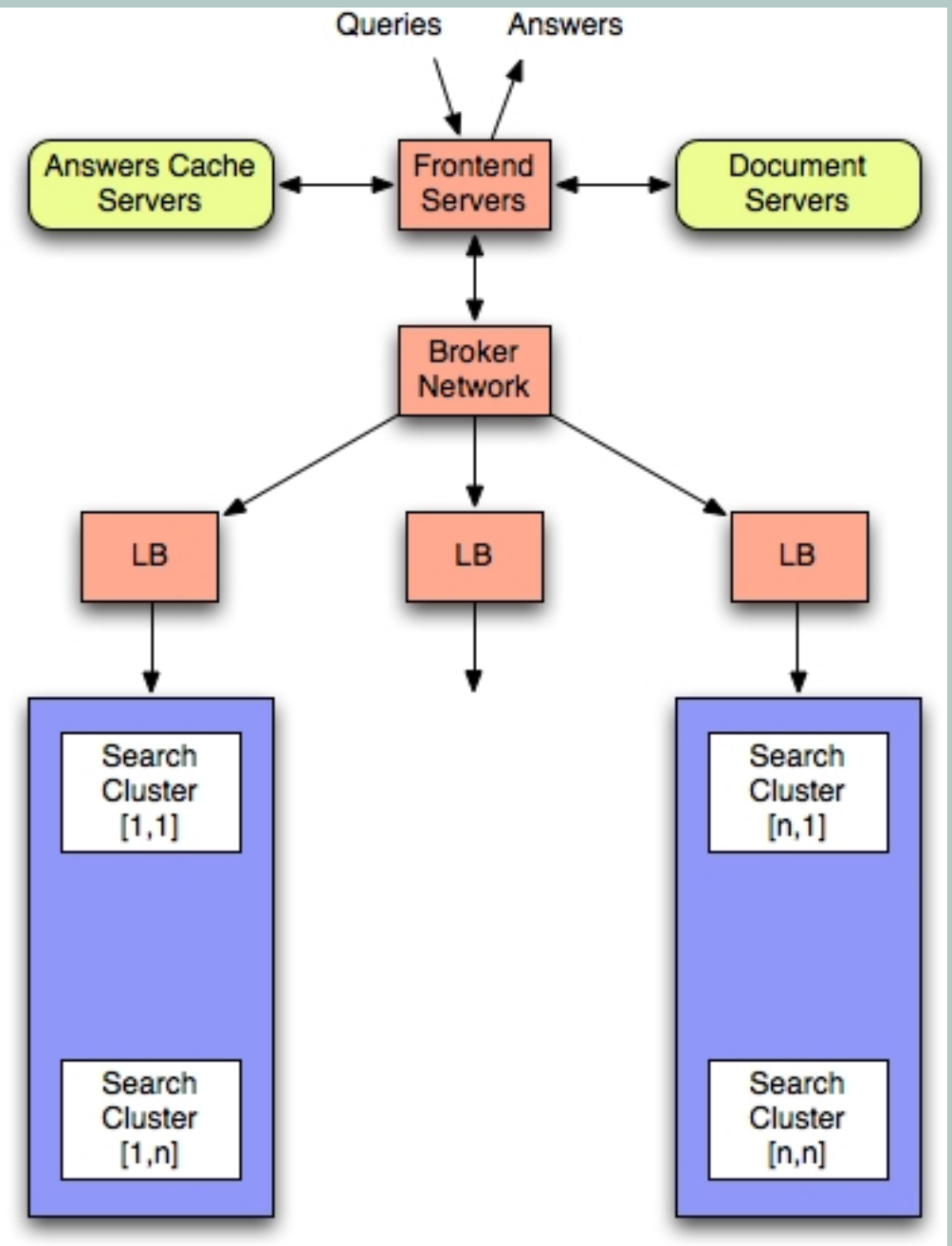
Arquitetura Típica de um Crawler



Arquitetura Típica de uma Engine de Busca [1]



Arquitetura baseada em Cluster para Busca



Consultando um Servidor Web

- Conexão a um servidor web utilizando o aplicativo *curl*
- Num prompt, digite o comando abaixo:
> *curl <url>*

Uma Taxonomia para Crawlers

- Questões a se considerar no projeto/implementação de um crawler
 - Atualização das páginas: páginas mais atualizadas possível x páginas “estáticas”
 - Qualidade: poucas páginas com muita qualidade x muitas páginas com diferentes níveis de qualidade
 - Quantidade: muitas páginas x maior atualização e/ou qualidade

O que é Web Mining?

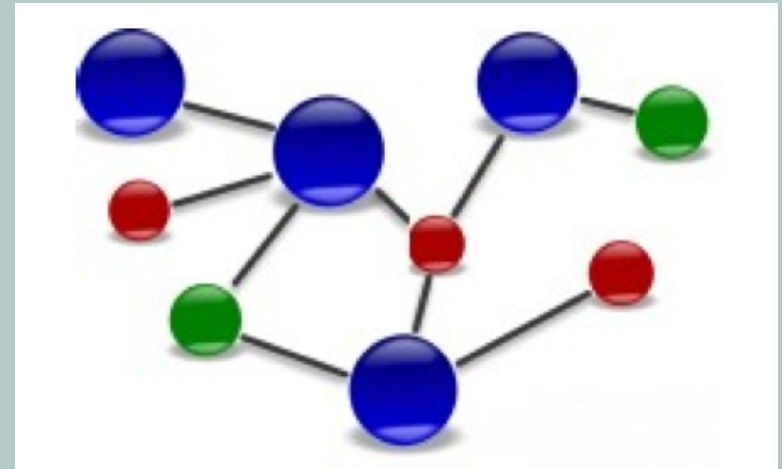
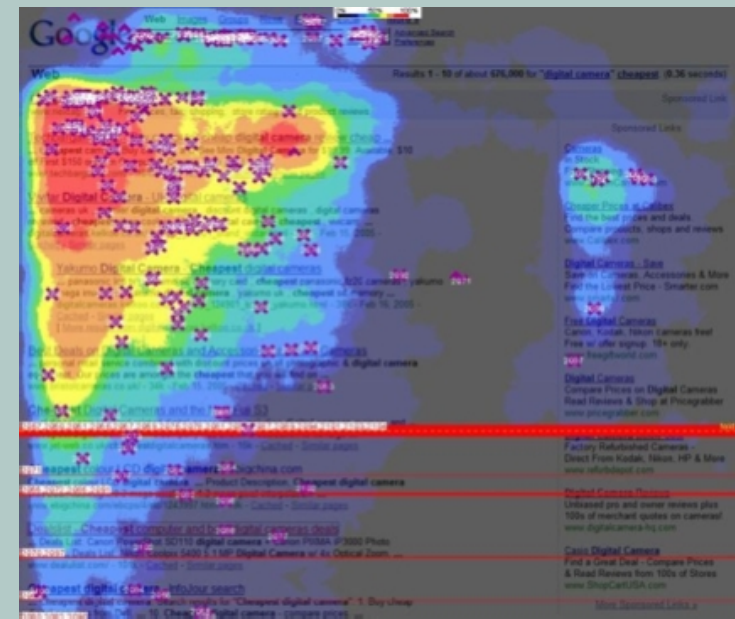
- Web Mining = Web + Data Mining
 - Information Retrieval, Machine Learning, Statistic, Pattern Recognition

		Data/information sources		
		Any data	Textual data	Web-related data
Purpose	Retrieving known data or documents efficiently and effectively	Data Retrieval	Information Retrieval	Web Retrieval
	Finding new patterns or knowledge previously unknown	Data Mining	Text Mining	Web Mining

O que é Web Mining?

- Fontes para Mineração na Web:
 - Conteúdo: textos, mídias, ...
 - Estrutura: links, âncoras, ...
 - Uso: navegação (“wisdom of crowds”)

Web Mining – Fontes

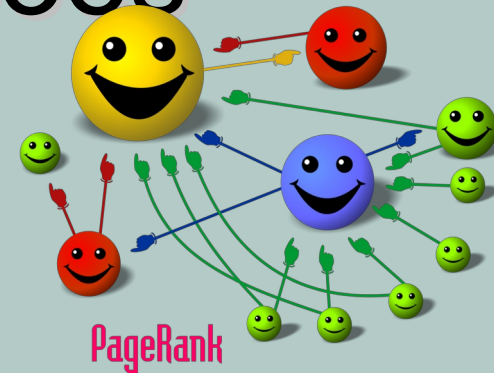
[illegible]

Exemplos de Aplicações

- PageRank

(Algoritmo de “ranqueamento”)

- Mineração na estrutura das páginas
- Uma página tem um bom pagerank se apontam para ela muitas outras
- Este valor aumenta se as páginas que apontam possuem um bom valor



Exemplos de Aplicações

- Google AdWords (Propaganda)
 - Mineração nas queries / conteúdo
 - Exibe conteúdo relacionado aos termos pesquisados
- Google Analytics
 - Mineração no conteúdo e navegação
 - Registra informações estatísticas de acesso e permanência num site

Exemplos de Aplicações

- Internet Archive (crawler de amplitude global) - <http://archive.org>
 - Projeto iniciado para armazenamento de versões de imagens de páginas web
 - Exemplos de uso: “www.nytimes.com” em 11/09/2001”, “www.cade.com.br”

Exemplos de Aplicações

- WolframAlpha (Engine para Consulta de Informações -
<http://www.wolframalpha.com>
 - Utiliza uma base de conhecimento para resposta às consultas
 - No site não descreve se há mineração para busca de informações adicionais
 - Exemplo de busca: “16h President of Brazil”, “Hebe Camargo birthdate”

Exemplos de Aplicações

- Netflix (Mineração no Uso)
 - Netflix Prize
(http://en.wikipedia.org/wiki/Netflix_Prize)
 - Algoritmos para Recomendação baseado em Visualização
 - Entrevista com funcionários da Netflix:
http://www.wired.com/underwire/2013/08/q_q_netflix-algorithm/

Exemplos de Aplicações

- NSA (Mineração ???)
 - U.S. National Security Agency
 - Imagens seguintes extraídas do site archive.org

Exemplos de Aplicações

TOP SECRET//SI//ORCON//NOFORN



Hotmail

YAHOO!

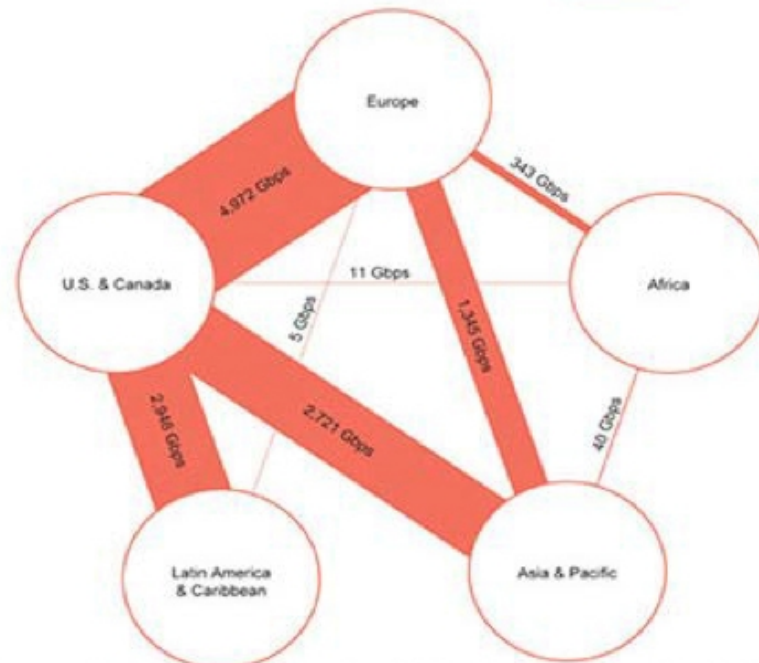


(TS//SI//NF) Introduction

U.S. as World's Telecommunications Backbone



- Much of the world's communications flow through the U.S.
- A target's phone call, e-mail or chat will take the **cheapest** path, **not the physically most direct** path – you can't always predict the path.
- Your target's communications could easily be flowing into and through the U.S.



International Internet Regional Bandwidth Capacity in 2011

Source: Telegeography Research

TOP SECRET//SI//ORCON//NOFORN

Exemplos de Aplicações

TOP SECRET//SI//ORCON//NOFORN

Hotmail® Google Apple Skype paltalk.com YouTube AOL mail

facebook

 (TS//SI//NF) PRISM Collection Details 

Current Providers

- Microsoft (Hotmail, etc.)
- Google
- Yahoo!
- Facebook
- PalTalk
- YouTube
- Skype
- AOL
- Apple

What Will You Receive in Collection (Surveillance and Stored Comms)?
It varies by provider. In general:

- E-mail
- Chat – video, voice
- Videos
- Photos
- Stored data
- VoIP
- File transfers
- Video Conferencing
- Notifications of target activity – logins, etc.
- Online Social Networking details
- **Special Requests**

Complete list and details on PRISM web page:
Go PRISMFAA

TOP SECRET//SI//ORCON//NOFORN

Exemplos de Aplicações

- Google Knowledge Graph
- Facebook Open Graph
- IBM Watson

Desafios na Análise de Dados na Web

- Dados distribuídos
- Dados voláteis
- Grande volume de dados
- Dados não estruturados e redundantes
- Qualidade dos dados
- Formatos heterogêneos
- Como expressar consultas
- Como interpretar os resultados

Referências

- [1] *Searching the Web*, Arvind Arasu et. al,
Journal ACM Transactions on Internet
Technology
- [2] Web Mining Research Survey,
<https://arxiv.org/pdf/cs/0011033.pdf>
- [3] Web Mining: Examples and Applications,
Arne Pottharst