

# Laboratório de Mídias Sociais

## Aula 04

Análise de rede de links utilizando crawlers

**Prof. Dalton Martins**

[dmartins@gmail.com](mailto:dmartins@gmail.com)

Gestão da Informação

Universidade Federal de Goiás




# Minerando os links de páginas web: fazendo crawler nas páginas

- Nas aulas passadas, aprendemos como utilizar ferramentas webométricas baseadas em mecanismos de buscas profissionais;
- Na aula de hoje, veremos como nós mesmos podemos coletar as páginas de um site, baixar seu conteúdo e analisar os links presentes nessas páginas;
- Esse recurso é sem dúvida mais sofisticado, porém é também computacionalmente mais oneroso e mais complexo de realizar;
- Vale utilizar para análise de sites pequenos e médios e para análise de redes webométricas com poucos sites (em torno de 100, no máximo) com esse perfil.

SocSciBot: Link crawler for the ...

socscibot.wlv.ac.uk

Pesquisar

 **Statistical Cybermetrics**  
Research Group

**SocSciBot**

› SocSciBot Home › Tutorial 1 › Tutorial 2 › Tutorial 3 › Tutorial 4 › Linguistics › FAQ › Link analysis book

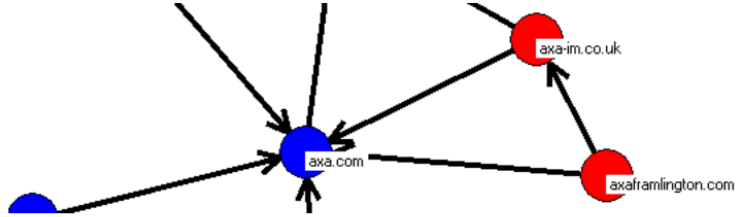
---


## SocSciBot

*Web crawler and link analyser for the social sciences and humanities*


**SocSciBot is a Web crawler for link analysis research** on a web site or collection of web sites, or for text search/analysis on a collection of sites. [Free SocSciBot download](#).

SocSciBot (a) crawls one or more web sites and (b) analyses them to produce standard statistics about their interlinking and network diagrams of the interlinking. It also runs limited analyses of the text in the web sites. To analyse links to one or more web sites, use [Webometric Analyst](#) instead. SocSciBot can export network diagrams to Pajek and to UCINET. See the [quick network tutorial](#).



 **SocSciBot and associated software: Conditions of use**

- › **Licence** SocSciBot is licenced free for non-commercial purposes only. We also do not accept liability for any damage resulting from its use, or for loss of data or other problems caused by the operations of the programs downloaded.
- › **Notification ethics** You must enter your email address into the program when requested, and check



## News

April 2016 general update.

February 2012 update allows quick partial crawling of multiple sites.

Instructions for [creating networks for large collections of web sites](#).

SocSciBot 4.1 has a button for [calculating link networks for the links between a set of web sites](#), plus improved network diagram functions.

[SocSciBot 4 blog](#) has more information and bug reports.

Windows taskbar: 21:10 08/09/2016

<http://socscibot.wlv.ac.uk/>

SocSciBot Link crawler for the ... Site do Laboratório de Polit... Site do Laboratório de Polit... +

socscibot.wlv.ac.uk/index.html#SocSciBotDownload

Pesquisar

Please fully collect your data with SocSciBot before opening the results with any other programs.

### SocSciBot Download Instructions

Download the programs only if you accept the conditions of use.

- › I have read the conditions of use (above) and accept them and now wish to *download SocSciBot 4*. **PLEASE READ TUTORIAL 1 BEFORE FIRST USE** as the the first use initialises some important parts of the software.

Tutorials and extra information for SocSciBot 4.

- › **Tutorial 1: Introduction to SocSciBot.** Please go through this immediately upon downloading SocSciBot.
- › **Tutorial 2: Mini link analysis research project case study** Using SocSciBot for a small scale link analysis research project - key features for link analysis research.
- › **Tutorial 3: Summary of how to use SocSciBot for a link analysis research project.**
- › **Corpus Linguistics Tutorial: Using SocSciBot for text analysis/basic corpus linguistics.** *SocSciBot includes concordancer/search engine interface, *Cyclist*, to the text of your downloaded sites.*
- › **Frequently Asked Questions.**
- › To convert large link structure files to Pajek, **Tobias Escher** of UCL has supplied a **Perl program**.

No technical support is provided, sorry.

The program runs on Windows only and will crawl sites with up to 15,000 pages and has a speed restriction. If you wish to crawl more pages or faster, please email your request. For example, we allow faster and bigger crawls of the university web sites of richer countries.

SocSciBot may not work well on web sites with non-ASCII URLs (e.g., Chinese) and the text analysis does not work well with non-ASCII text.

There is an article describing the database structure and crawler linked from the [cybermetrics database site](#). Please ignore all the numbers reported by the program, both in its title bar and in the summary file produced - these are for testing purposes. The reliable information is in the link data file and the text data file, but you may need to use the cybermetrics programs to get at this information.

SocSciBot can be used on its own or in conjunction with the [link analysis book](#) or the [Introduction to Webometrics](#) book.

PT 21:22 08/09/2016

Veja as condições de uso e baixe o programa...



**Wizard Step 0 - Data location**

This seems to be your first use of this program. You need to select a folder on your computer to save all of your crawl data to. If the folder below is not OK, please change to a different one.

Suggested folder:

WHENEVER YOU CRAWL A SITE, SocSciBot will EMAIL THE WEBMASTER your email address to notify them. You must enter your email address below and check your email when you crawl. If the webmaster asks you not to crawl their site, you must stop the crawl straight away. \*\* IT IS A CONDITION OF USING THIS PROGRAM THAT YOU ENTER YOUR CORRECT EMAIL ADDRESS. \*\*

Your email address:

If you wish to describe the purpose of your crawls, please enter a description below. This will be emailed to webmasters of sites that you crawl, in addition to the standard information given, such as your email address and the fact that you are about to crawl the site.

Your message (optional):

Defina o diretório onde os arquivos serão armazenados!

Wizard Step 1

Existing Projects

Click on a project to select it - or type in a new project name below

To start a new project, enter name and click the 'Start new project' button

Exit

Go to graph drawing program

Not using proxy Cache

SocSciBot 4 version 2.0.6005.14298 restricted mode  
[SocSciBot Online Tutorials and Help](#)


Vamos criar um novo projeto para avaliar os links entre os sites de cultura que vimos nas aulas anteriores.




Wizard Step 2: Project "cultura"


**1.** STEP 2a. Crawl your sites individually or simultaneously.  
Crawl sites individually if they are big or if you want to do a text analysis.

☒ Download multiple sites/URLs in one crawl (not for text analysis)

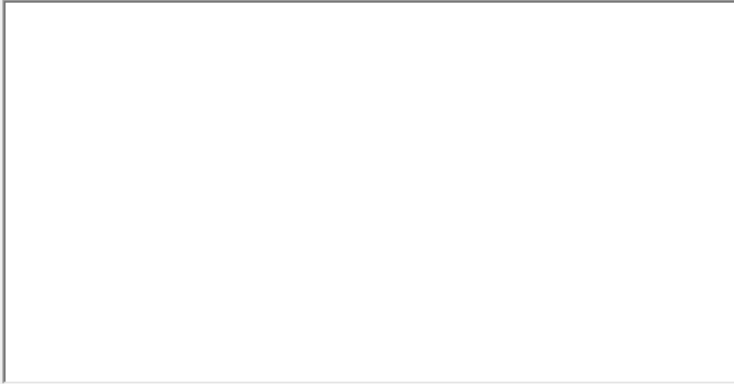
 Crawl Site(s)  
with SocSciBot

**2.** STEP 2b. Analyse web sites when all crawls are finished. [You can add crawls to the project after. Do not analyse a project during a crawl.]

 Analyse LINKS  
with SocSciBot Tools

 Analyse TEXT  
with Cyclist

Completed crawls - Click to delete



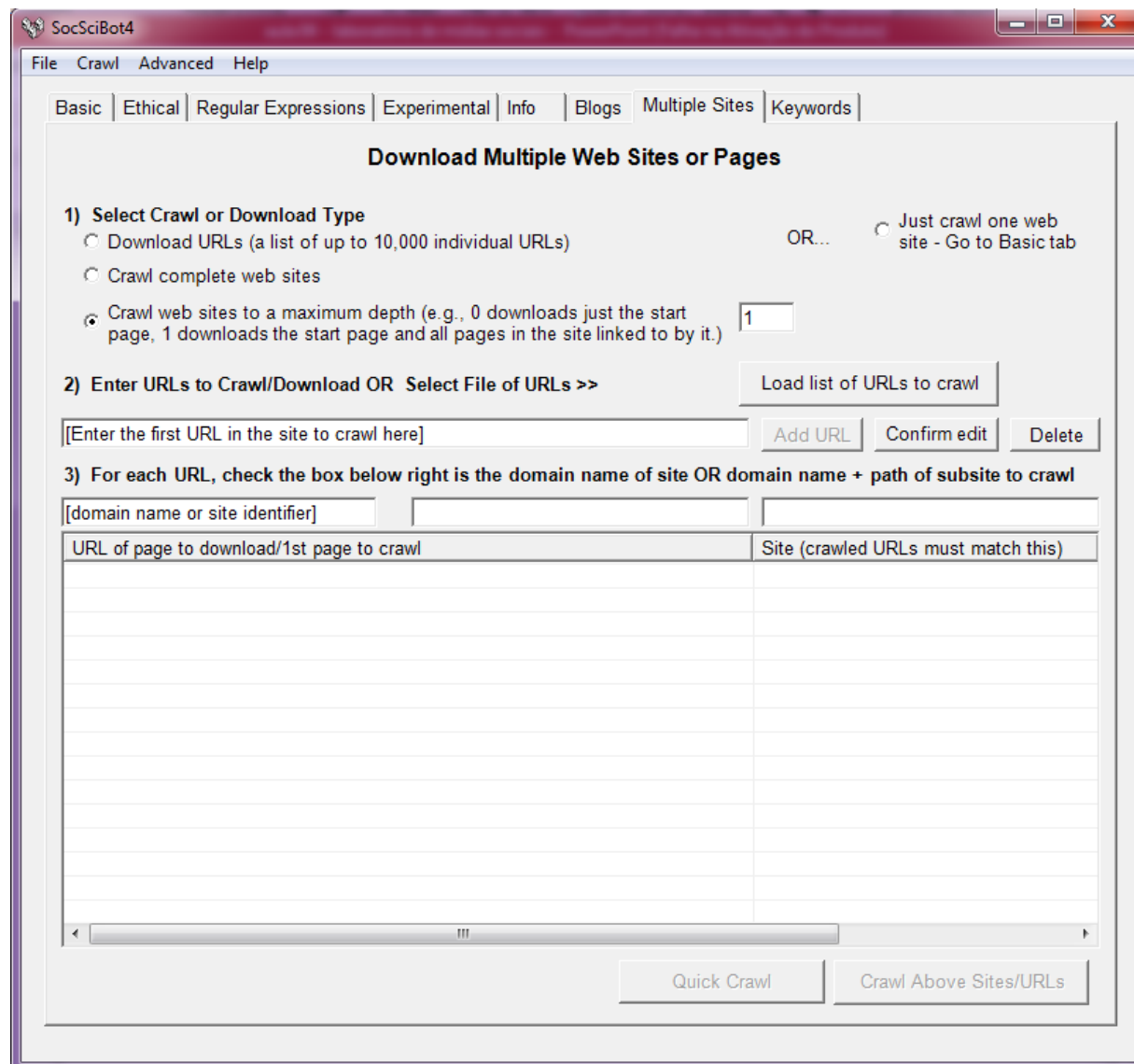
**\*\*Warning: deleted crawls cannot be recovered!\*\***

[SocSciBot Online Tutorials and Help](#)

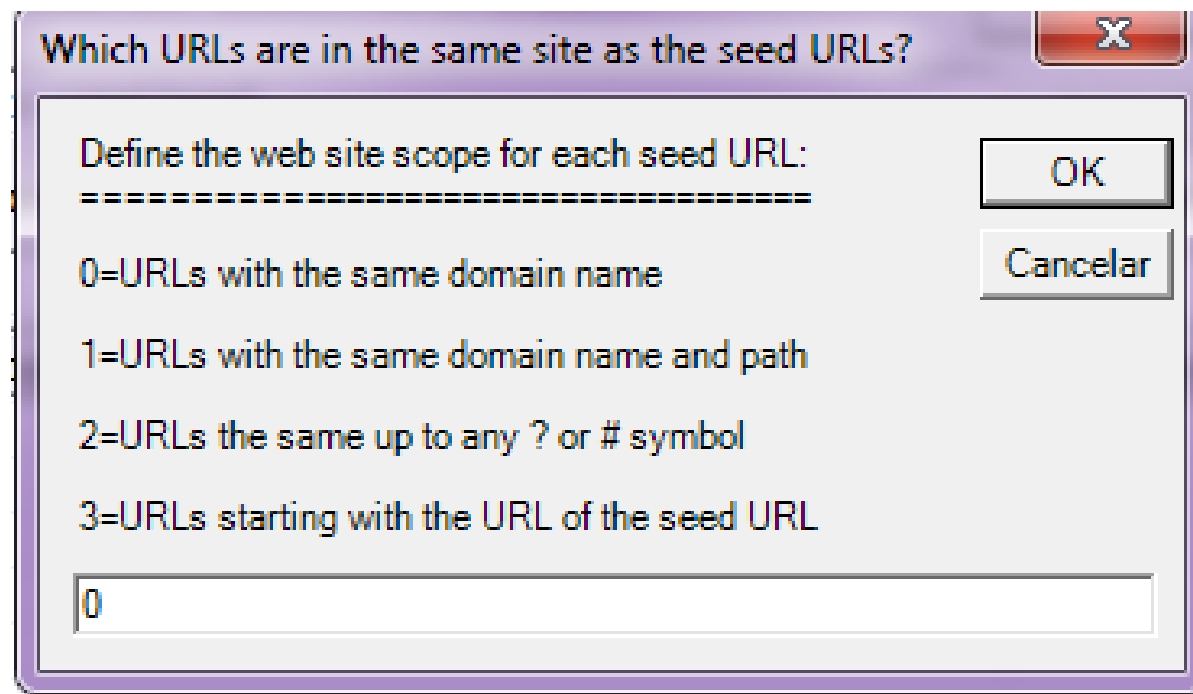
Choose different project      Exit

Selecionamos a opção para download de múltiplos sites e vamos pra próxima tela para carregar os sites...



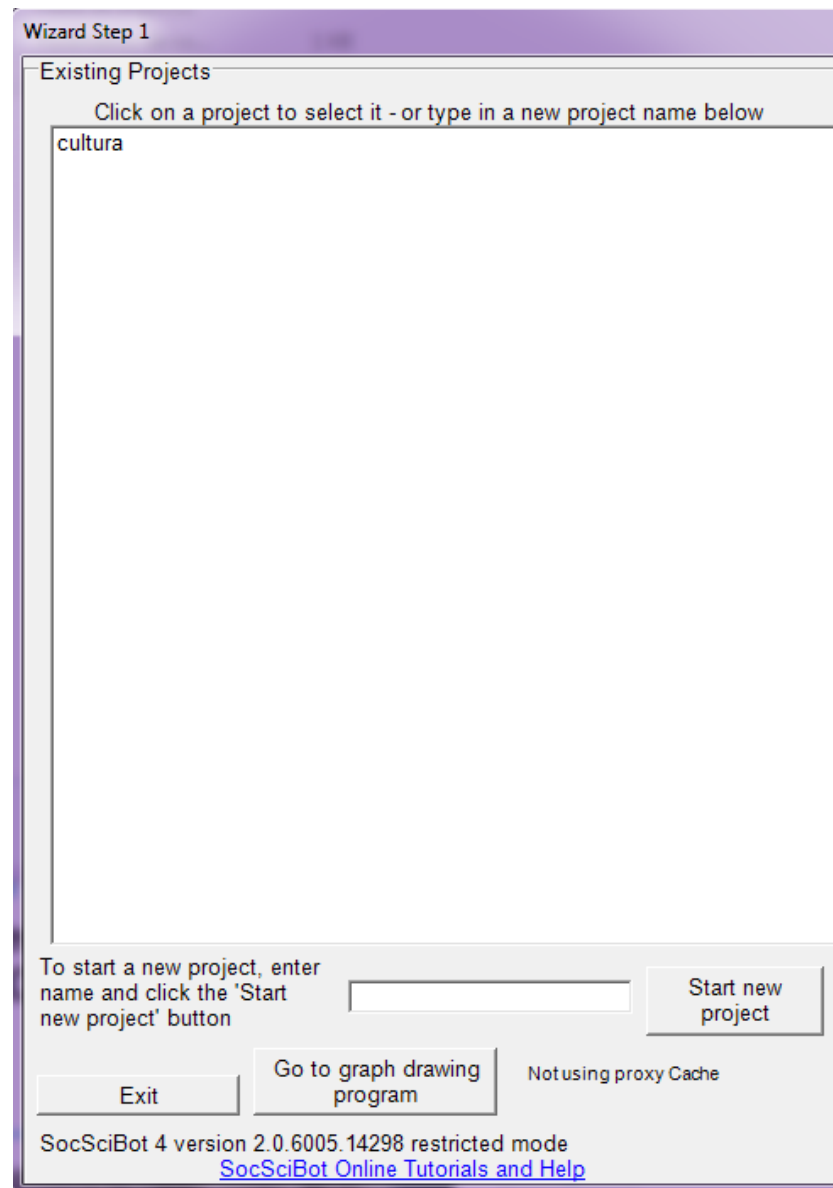


Vamos configurar para baixar apenas a pagina home e as paginas linkadas a ela. Daí, vamos carregar a lista de páginas para análise.



Vamos configurar apenas para coletar as páginas que tiverem o mesmo nome do domínio.





Selecionamos o projeto criado e vamos direto para a pagina de análises da ferramenta.


Wizard Step 2: Project "cultura"

**1.** STEP 2a. Crawl your sites individually or simultaneously.  
Crawl sites individually if they are big or if you want to do a text analysis.


**Enter site home page or other starting URL:**


**OR**

☐ Download multiple sites/URLs in one crawl (not for text analysis)

 Crawl Site(s)  
with SocSciBot

**2.** STEP 2b. Analyse web sites when all crawls are finished. [You can add crawls to the project after. Do not analyse a project during a crawl.]

 Analyse LINKS  
with SocSciBot Tools

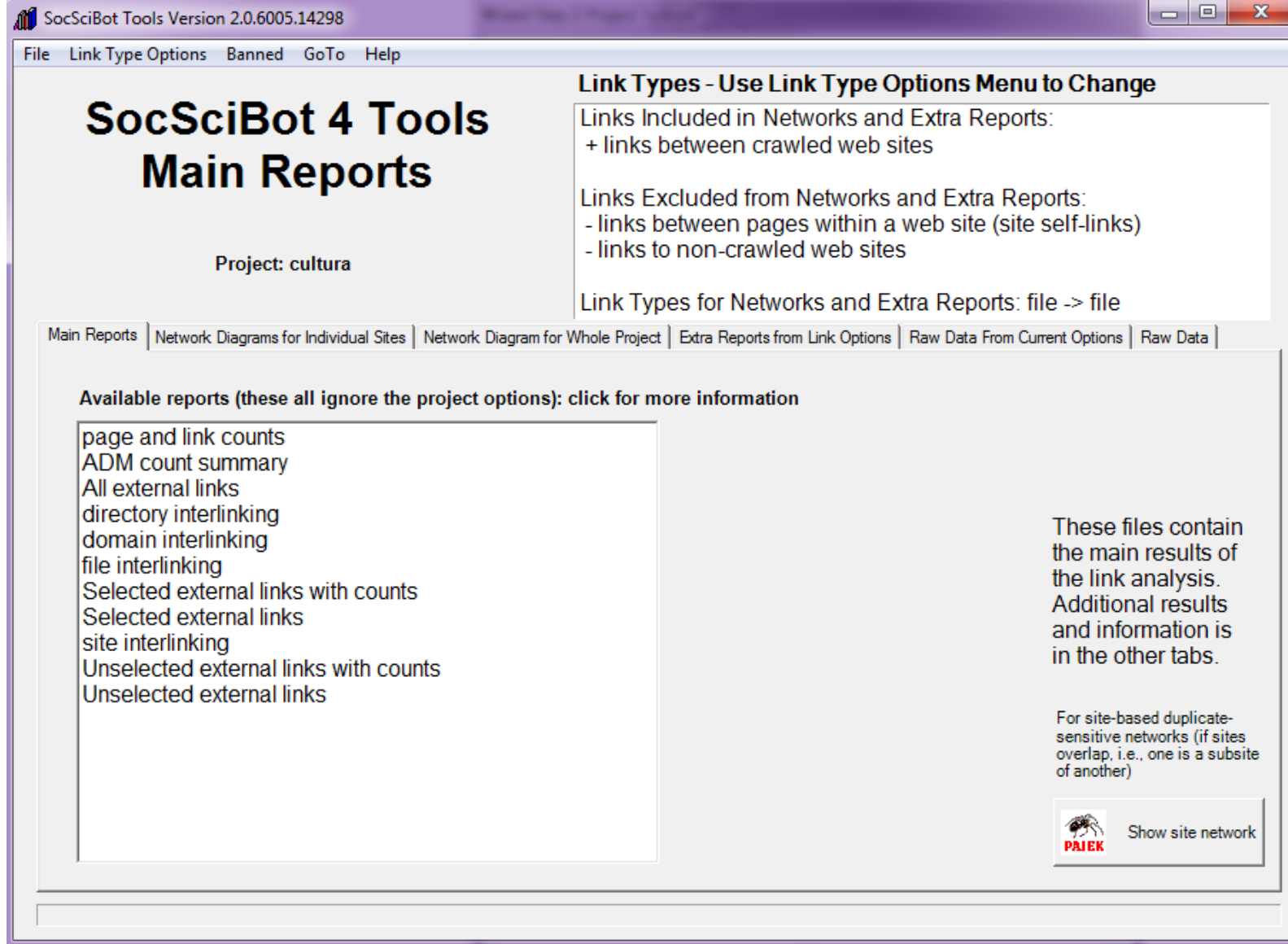
 Analyse TEXT  
with Cyclist

Completed crawls - Click to delete

**\*\*Warning: deleted crawls cannot be recovered!\*\***

[SocSciBot Online Tutorials and Help](#)

Escolhemos o botão “Analyse LINKS with SocSciBot Tools”.



Temos varias opções de relatórios a partir desses sites. Vamos explorar alguns. Vale notar que ao clicar em cada opção, se abre uma pequena janela que mostra o que significa aquele resultado e algumas formas de visualizar os dados em formato texto, excel e visualização de redes.

SocSciBot Tools Version 2.0.6005.14298

File Link Type Options Banned GoTo Help

# SocSciBot 4 Tools Main Reports

Project: cultura

## Link Types - Use Link Type Options Menu to Change

Links Included in Networks and Extra Reports:  
+ links between crawled web sites


Links Excluded from Networks and Extra Reports:  
- links between pages within a web site (site self-links)  
- links to non-crawled web sites


Link Types for Networks and Extra Reports: file -> file

Main Reports | Network Diagrams for Individual Sites | Network Diagram for Whole Project | Extra Reports from Link Options | Raw Data From Current Options | Raw Data

Available reports (these all ignore the project options): click for more information

- page and link counts
- ADM count summary
- All external links
- directory interlinking
- domain interlinking
- file interlinking
- Selected external links with counts
- Selected external links
- site interlinking**
- Unselected external links with counts
- Unselected external links

 View report

 View in Excel

### site interlinking

=====


This reports the number of links between pairs of web sites in different sites (one link per pair of web sites).

=====

Each line of the file contains:  
source -tab- target -tab- counts,  
which looks ugly in Notepad but  
can be copied or imported into a  
spreadsheet or database program.

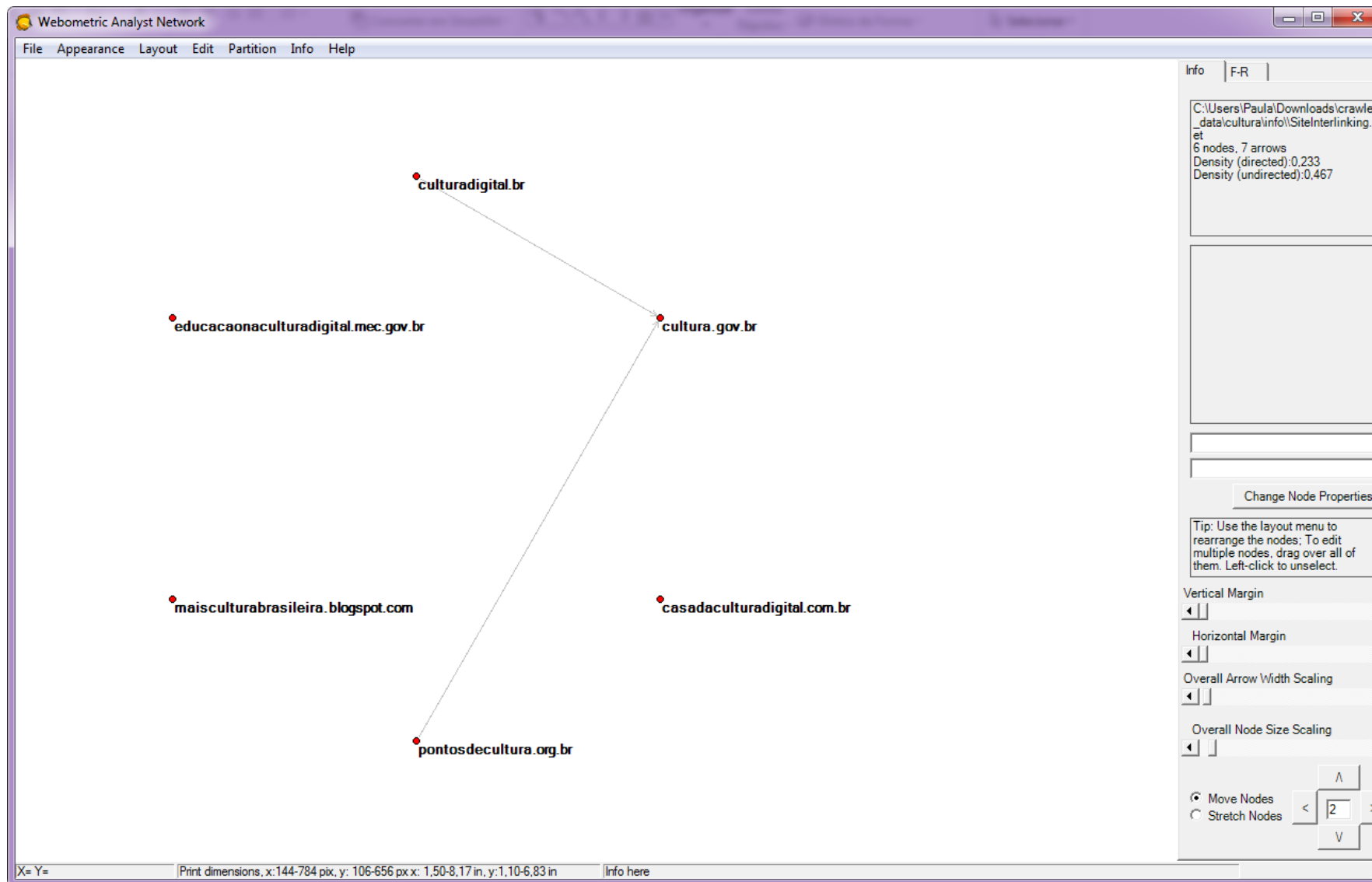
These files contain the main results of the link analysis. Additional results and information is in the other tabs.

For site-based duplicate-sensitive networks (if sites overlap, i.e., one is a subsite of another)

 Show site network

Vamos analisar a interconexão entre os sites pelo relatório “site interlinking”.





Entramos na ferramenta de análise de redes, como do Webometric Analyst.  
Vamos comparar com o resultado anterior!!! Vejamos o arquivo da aula passada....

Vamos agora ver a analise de apenas um site...

Vamos usar as mesmas telas, mas com parâmetros diferentes.