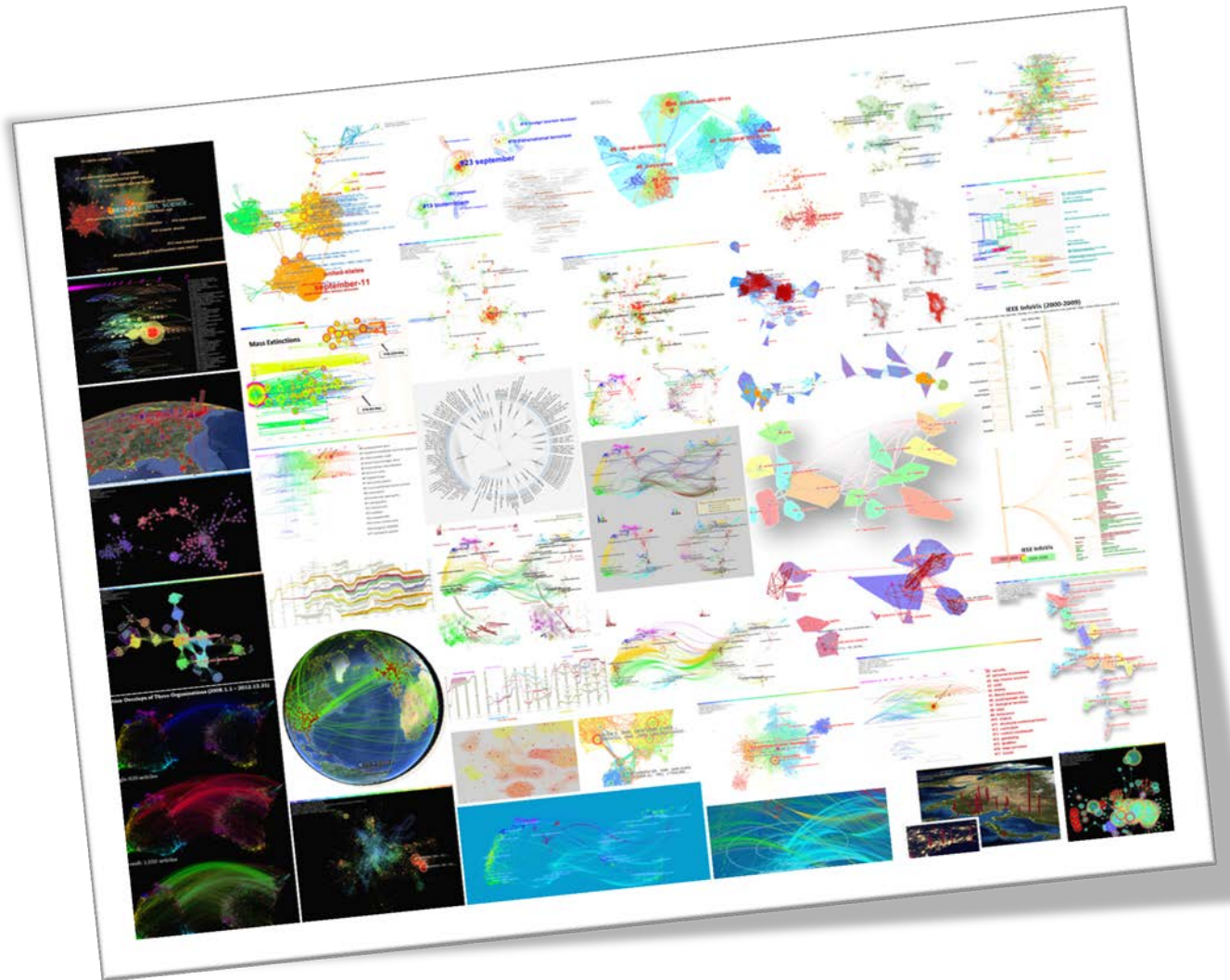


# The *CiteSpace* Manual

Version 1.01

Chaomei Chen  
College of Computing and Informatics  
Drexel University



How to cite:

Chen, Chaomei (2014) The *CiteSpace* Manual. <http://cluster.ischool.drexel.edu/~cchen/citespace/CiteSpaceManual.pdf>

## Contents

1	How can I find the latest version of the CiteSpace Manual? .....	5
2	What can I use CiteSpace for? .....	5
2.1	What if I have Questions .....	7
2.2	How should I cite CiteSpace? .....	7
2.3	Where are the Users of CiteSpace? .....	8
3	Requirements to Run CiteSpace .....	10
3.1	Java Runtime (JRE).....	10
3.2	How do I check whether Java is on my computer?.....	10
3.3	Do I have a 32-bit or 64-bit Computer? .....	12
4	How to Install and Configure <i>CiteSpace</i> .....	12
4.1	Where Can I download CiteSpace from the Web? .....	12
4.2	What is the maximum number of records that I can handle with CiteSpace? .....	13
4.3	How to configure the memory allocation for CiteSpace? .....	13
4.4	How to uninstall CiteSpace .....	14
4.5	On Mac or Unix-based Systems.....	15
5	Get Started with CiteSpace .....	19
5.1	Try it with a demonstrative dataset .....	19
5.1.1	The Demo Project .....	20
5.1.2	Clustering.....	23
5.1.3	Generate Cluster Labels.....	25
5.1.4	Where are the major areas of research based on the input dataset?.....	27
5.1.5	How are these major areas connected? .....	28
5.1.6	Where are the most active areas?.....	28
5.1.7	What is each major area about? Which/where are the key papers for a given area? 36	
5.1.8	Timeline View .....	38
5.2	Try it with a dataset of your own .....	39
5.2.1	Collecting Data .....	39
5.2.2	Working with a CiteSpace Project.....	43
5.2.3	Data Sources in Chinese .....	44
5.2.4	How to handle search results containing irrelevant topics.....	45
6	Configure a CiteSpace Run.....	47
6.1	Time Slicing .....	47

6.2	Text Processing .....	48
6.3	Configure the Networks .....	48
6.3.1	Bibliographic Coupling.....	49
6.4	Node Selection Criteria .....	49
6.4.1	Do I have the right network? .....	50
6.5	Pruning, or Link Reduction .....	50
6.6	Visualization.....	51
7	Interacting with CiteSpace .....	51
7.1	How to Show or Hide Link Strengths .....	51
7.2	Adding a Persistent Label to a Node .....	52
7.3	Using Aliases to Merge Nodes.....	53
7.4	How to Exclude a Node from the Network.....	55
7.5	How to Use the Fisheye View Slider .....	55
7.6	How to Configure When to Calculate Centrality Scores Automatically .....	56
7.7	How to Save the Visualization as a PNG File.....	57
7.8	Filters: Match Records with Pubmed.....	58
8	Additional Functions.....	62
8.1	Menu: Data.....	62
8.1.1	CiteSpace Built-in Database .....	62
8.1.2	Utility Functions for the Web of Science Format.....	65
8.1.3	Scopus .....	66
8.1.4	PubMed.....	67
8.2	Menu: Network .....	69
8.2.1	Batch Export to Pajek .net Files.....	69
8.3	Menu: Geographical.....	69
8.3.1	Generate Google Earth Maps.....	69
8.4	Menu: Overlay Maps.....	72
8.4.1	Add an Overlay .....	73
8.4.2	Further Reading and Terms of Use.....	75
8.5	Menu: Text.....	75
8.5.1	Concept Trees and Predicate Trees.....	75
8.5.2	List Terms by Clumping Properties.....	78
8.5.3	Latent Semantic Analysis .....	79
9	Selected Examples .....	80

10	Metrics and Indicators.....	82
10.1	Information Theoretic.....	82
10.1.1	Information Entropy.....	82
10.2	Structural .....	82
10.2.1	Betweenness Centrality.....	82
10.2.2	Modularity.....	82
10.2.3	Silhouette .....	82
10.3	Temporal.....	82
10.3.1	Burstness .....	82
10.4	Combined.....	82
10.4.1	Sigma .....	82
10.5	Cluster Labeling .....	83
10.5.1	Term Frequency by Inversed Document Frequency.....	83
10.5.2	Log-Likelihood Ratio.....	83
10.5.3	Mutual Information.....	83
11	References.....	83

## 1 How can I find the latest version of the CiteSpace Manual?

The latest version of the CiteSpace Manual is always at the following location:

<http://cluster.ischool.drexel.edu/~cchen/citespace/CiteSpaceManual.pdf>

You can also access the manual from CiteSpace: Help ► View the CiteSpace Manual (PDF). It will open up the PDF file in a new browser window.

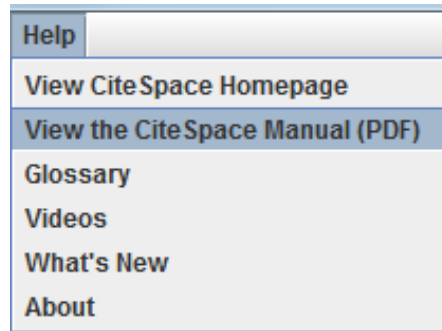


Figure 1. The latest version of the CiteSpace Manual is accessible from CiteSpace itself.

## 2 What can I use CiteSpace for?

CiteSpace is designed to answer questions about a knowledge domain, which is a broadly defined concept that covers a scientific field, a research area, or a scientific discipline. A knowledge domain is typically represented by a set of bibliographic records of relevant publications. It is your responsibility to prepare the most appropriate and representative dataset that contains adequate information to answer your questions.

CiteSpace is designed to make it easy for you to answer questions about the structure and dynamics of a knowledge domain. Here are some typical questions:

- What are the major areas of research based on the input dataset?
- How are these major areas connected, i.e. through which specific articles?
- Where are the most active areas?
- What is each major area about? Which/where are the key papers for a given area?
- Are there critical transitions in the history of the development of the field? Where are the 'turning points'?

The design of CiteSpace is inspired by Thomas Kuhn's structure of scientific revolutions. The central idea is that centers of research focus change over time, sometime incrementally and other times drastically. The development of science can be traced by studying their footprints revealed by scholarly publications.

Members of the contemporary scientific community make their contributions. Their contributions form a dynamic and self-organizing system of knowledge. The system contains consensus, disputes, uncertainties, hypotheses, mysteries, unsolved problems, and unanswered questions. It is not enough to study a single school of thought. In fact, a better understanding of a specific topic often relies on an understanding of how it is related to other topics.



## 2.1 What if I have Questions

If you have a question regarding the use of CiteSpace, you should first check the manual whether your question is answered in the manual. You can do a simple search through the PDF file to find out.

If the manual does not get you anywhere, you can ask your questions on the Facebook page of CiteSpace:

<https://www.facebook.com/pages/CiteSpace/276625072366558>

You can also post questions to my blog on sciencenet:

<http://blog.sciencenet.cn/home.php?mod=space&uid=496649>

Please refrain from sending me emails because you will have a much better chance to get my response from either the Facebook or the sciencenet blog.

Generally speaking, thoughtful questions get answered quickly. Questions that you may be able to figure out the answer for yourself if you think a little bit more about it would have a lower priority in the answering queue; it is quite possible that some of them never get answered.

## 2.2 How should I cite CiteSpace?

The following three publications represent the core ideas of CiteSpace.

The 2004 PNAS paper is the initial publication on CiteSpace (Chen 2004). In hindsight, it could have been named CiteSpace I. The 19-page 2006 JASIST paper gives the most thorough and in-depth description of CiteSpace II's key functions (C. M. Chen, 2006), plus a follow-up study of domain experts identified in the visualizations. The 2010 JASIST paper is even longer with 24 pages (C. Chen, Ibekwe-SanJuan, & Hou, 2010), which is the third of the trilogy. It describes technical details on how cluster labels are selected and how each of the three selection algorithms in comparison with labels chosen by domain experts.

Citations (Google Scholar)	Reference
800	Chen, C. (2006). "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature." <i>Journal of the American Society for Information Science and Technology</i> <b>57</b> (3): 359-377.
394	Chen, C. (2004). "Searching for intellectual turning points: Progressive Knowledge Domain Visualization." <i>Proc. Natl. Acad. Sci. USA</i> <b>101</b> (Suppl.): 5303-5310.
157	Chen, C., et al. (2010). "The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis." <i>Journal of the American Society for Information Science and Technology</i> <b>61</b> (7): 1386-1409.

The most recent case study of a topic outside the realm of information science and scientometrics is a scientometric study of regenerative medicine (C. Chen, Hu, Liu, & Tseng, 2012).

Chen, C., et al. (2012). "Emerging trends in regenerative medicine: A scientometric analysis in CiteSpace." *Expert Opinions on Biological Therapy* **12**(5): 593-608.

### 2.3 Where are the Users of CiteSpace?

In terms of the cities where CiteSpace were used, China, the United States, and Europe are prominent. Brazil, Turkey, and Spain also have many cities on the chart.

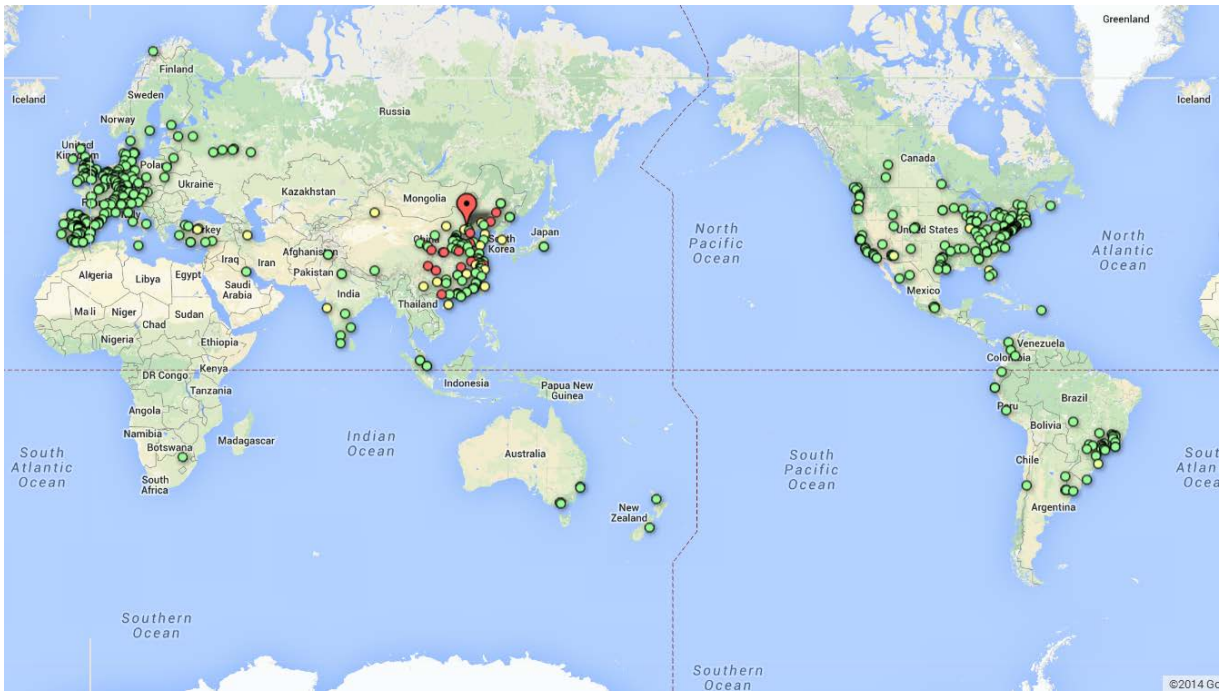


Figure 3. Cities with users of CiteSpace between August 2013 and March 2014 are shown on the map. The colors of markers depict the level of user intensity: green (1-10), yellow (10-100), red (100-1000), and the large red water drop shaped marker (1000+).

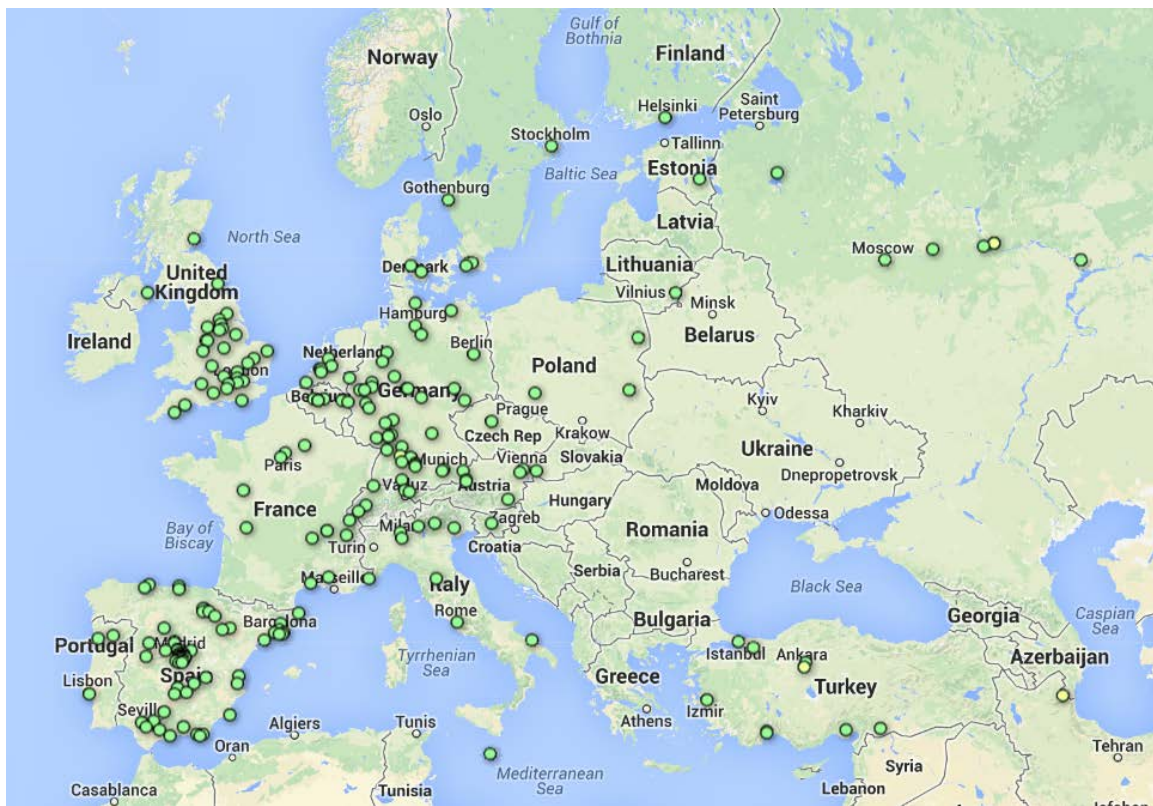


Figure 4. The use of CiteSpace in China (August 2013 – March 2014).





**Figure 5. The use of CiteSpace in the United States (August 2013 – March 2014).**



**Figure 6. The use of CiteSpace in Europe (August 2013 – March 2014).**

### 3 Requirements to Run CiteSpace

#### 3.1 Java Runtime (JRE)

CiteSpace is written in Java. It is a Java application. You should be able to run it on a computer that supports Java, including Windows or Mac.

CiteSpace is currently optimized for Windows 64-bit Java 7 (i.e. Java 1.7).

To run a Java application on your computer, you need to have Java Runtime (JRE) installed on your computer.

#### 3.2 How do I check whether Java is on my computer?

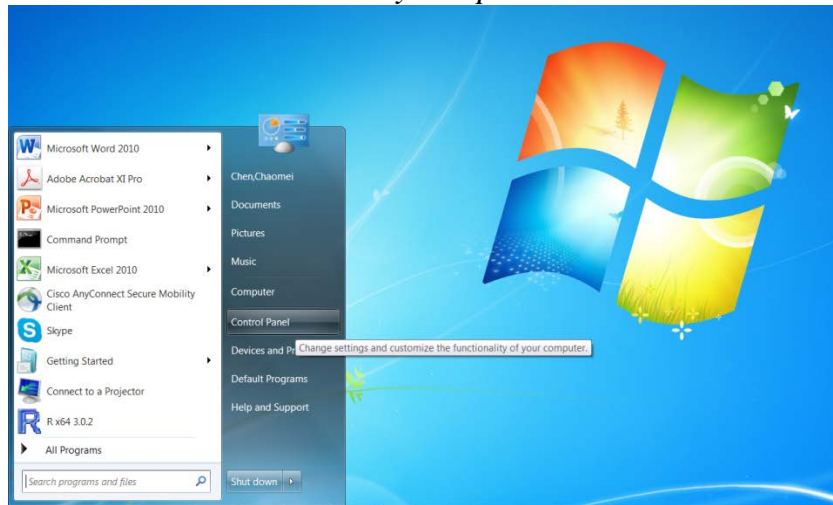
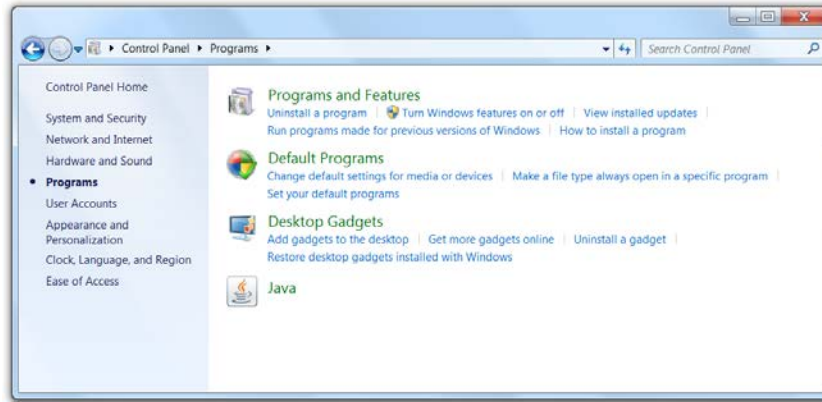


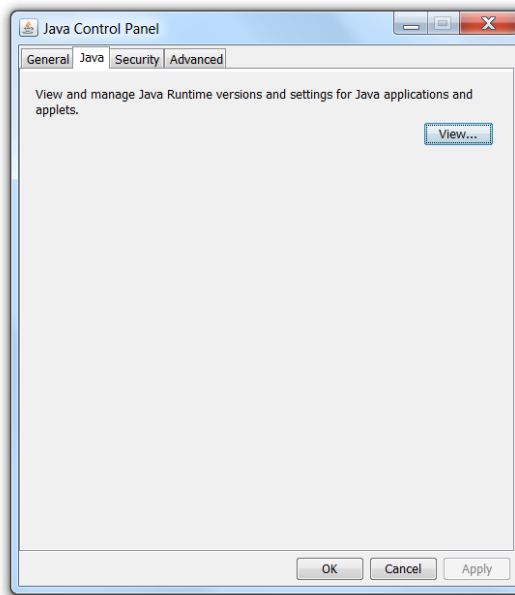
Figure 7. Select Control Panel.



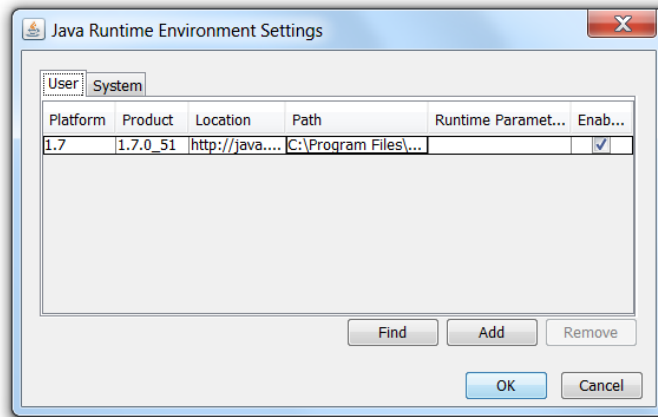
Figure 8. Click into the Programs category to find the Java control panel.



**Figure 9. Locate the Java control panel.**



**Figure 10. Java Control Panel. Choose the Java tab and press the View button to see more detail.**



**Figure 11. Java Runtime 1.7 is installed.**

### 3.3 Do I have a 32-bit or 64-bit Computer?

You need to find out whether your computer has a 32-bit or a 64-bit operating system.

Go to **Control Panel ► System and Security ► System**. You will see various details about your computer. Under the System type, you will see whether you have a 32-bit or a 64-bit operating system.

Follow the link below for further instructions on how to install Java:

[http://www.java.com/en/download/help/index\\_installing.xml](http://www.java.com/en/download/help/index_installing.xml)

Once you have Java Runtime setup on your computer, you can proceed to install CiteSpace.

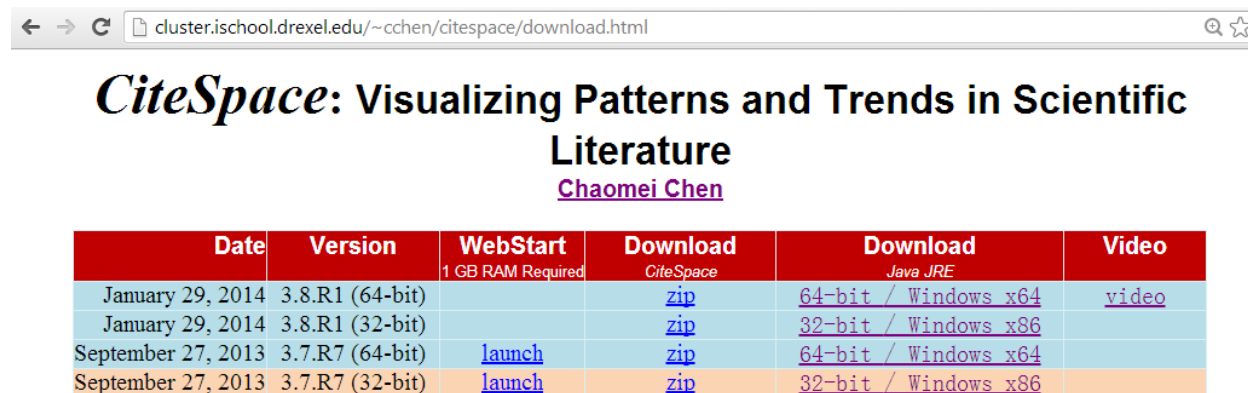
## 4 How to Install and Configure *CiteSpace*

CiteSpace is provided as a zip file for 64-bit and 32-bit computers. For Mac users, you need to download the 64-bit version.

### 4.1 Where Can I download CiteSpace from the Web?

You can download the latest version of CiteSpace from the following website:

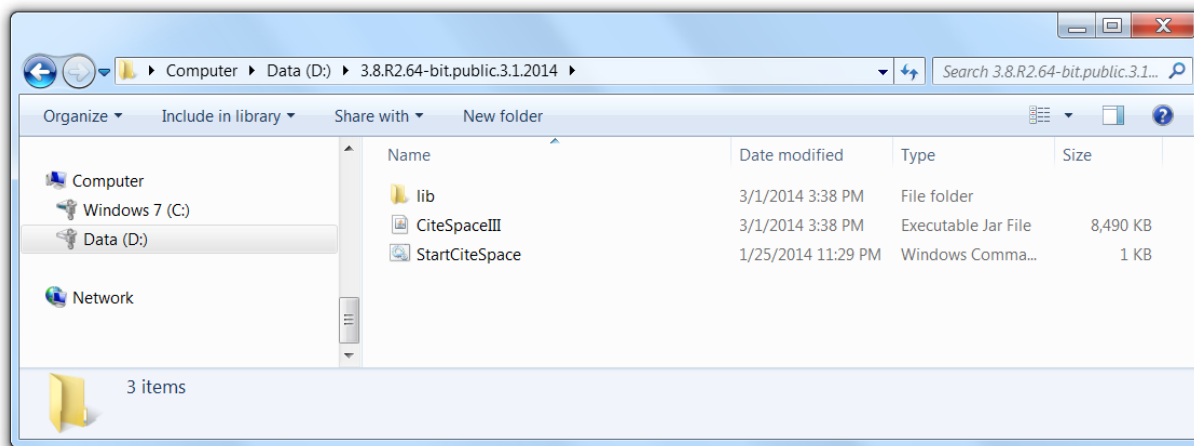
<http://cluster.ischool.drexel.edu/~cchen/citespace/download.html>



Date	Version	WebStart <small>1 GB RAM Required</small>	Download <small>CiteSpace</small>	Download <small>Java JRE</small>	Video
January 29, 2014	3.8.R1 (64-bit)		<a href="#">zip</a>	<a href="#">64-bit / Windows x64</a>	<a href="#">video</a>
January 29, 2014	3.8.R1 (32-bit)		<a href="#">zip</a>	<a href="#">32-bit / Windows x86</a>	
September 27, 2013	3.7.R7 (64-bit)	<a href="#">launch</a>	<a href="#">zip</a>	<a href="#">64-bit / Windows x64</a>	
September 27, 2013	3.7.R7 (32-bit)	<a href="#">launch</a>	<a href="#">zip</a>	<a href="#">32-bit / Windows x86</a>	

Figure 12. The download page of CiteSpace.

After you download the zip file to your computer, unpack the zip file to a folder of your choice.



**Figure 13.** CiteSpace is unpacked to the D drive on a computer.

Now you can start CiteSpace by double clicking on the StartCiteSpace file.

If you need to modify the amount memory allocated for CiteSpace (more precisely for Java Virtual Machine on which CiteSpace to be running), you can edit StartCiteSpace as a plain text file with any text editor.

#### 4.2 *What is the maximum number of records that I can handle with CiteSpace?*

This question needs to be answered at two levels: the number of records processed by CiteSpace and the number of nodes visualized, i.e. you can see and interact with them in CiteSpace.

The first number is the total number of records in your downloaded dataset. CiteSpace reads through each record in your download files.

The second number is determined by the selection criteria you specify and by the amount of memory, i.e. RAM, available on your computer. The more RAM you can make available for CiteSpace, the larger sized network you can visualize with a faster response rate.

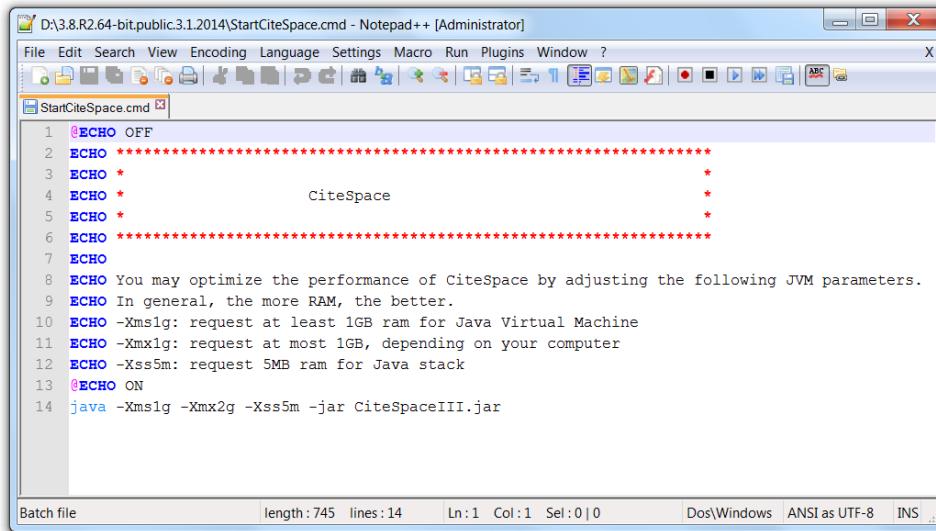
The speed of processing is also affected by a few computationally expensive algorithms such as Pathfinder network scaling and cluster labeling. Empirically, the best options for Pathfinder network scaling would be 50~500 nodes per slice. With faster computers or if you can wait for a bit longer, you can raise the number accordingly.

The completion time of cluster labeling is related to the size of your dataset. If the entire timespan of your dataset is 100 years but you will only need to consider the most recent 10 years, it will be a good idea to carve out a much smaller dataset as long as it covers the 10 years of interest. It will reduce the processing time considerably.

#### 4.3 *How to configure the memory allocation for CiteSpace?*

The performance of CiteSpace is influenced by the amount of memory accessible to the Java Virtual Machine (JVM) on which CiteSpace is running. To analyze a large amount of records, you should consider allocating as much as memory for CiteSpace to use.

You can modify the StartCiteSpace.cmd file to optimize the setting. More specifically, modify line 14 in the file. For example, `-Xmx2g` means that CiteSpace may get a maximum of 2GB of RAM to work with. Save the file after making any changes. And restart CiteSpace.

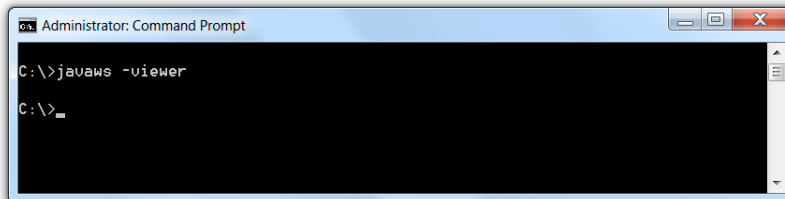


```
1 @ECHO OFF
2 ECHO *****
3 ECHO *
4 ECHO *           CiteSpace
5 ECHO *
6 ECHO *****
7 ECHO
8 ECHO You may optimize the performance of CiteSpace by adjusting the following JVM parameters.
9 ECHO In general, the more RAM, the better.
10 ECHO -Xms1g: request at least 1GB ram for Java Virtual Machine
11 ECHO -Xmx1g: request at most 1GB, depending on your computer
12 ECHO -Xss5m: request 5MB ram for Java stack
13 @ECHO ON
14 java -Xms1g -Xmx2g -Xss5m -jar CiteSpaceIII.jar
```

Figure 14. Configure the memory for Java in line 14.

#### 4.4 How to uninstall CiteSpace

You can use the following steps to remove cached copies of CiteSpace from your computer.



```
Administrator: Command Prompt
C:\>javaws -viewer
C:\>_
```

Figure 15. In a Command Prompt window, type `javaws -viewer`.

When you see a list of cached copies of CiteSpace in the Java Cache Viewer, select the items that you want to remove and then click on the button with a red cross.

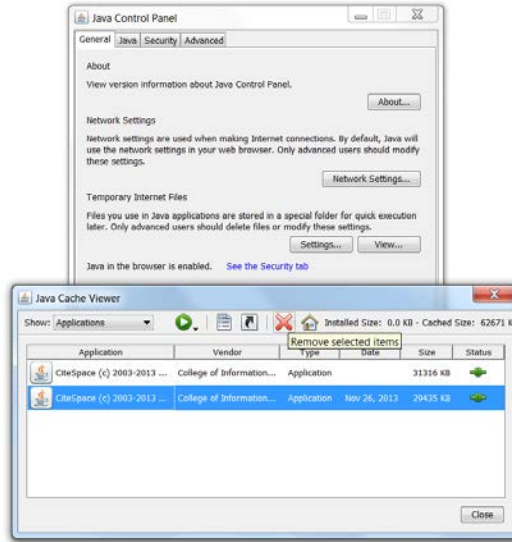


Figure 16. Select a cached copy of CiteSpace and remove the item.

#### 4.5 On Mac or Unix-based Systems

The following example shows you the basic steps to get started with CiteSpace on a Mac. First, go to the CiteSpace homepage in a browser such as Chrome and download the latest 64-bit version.

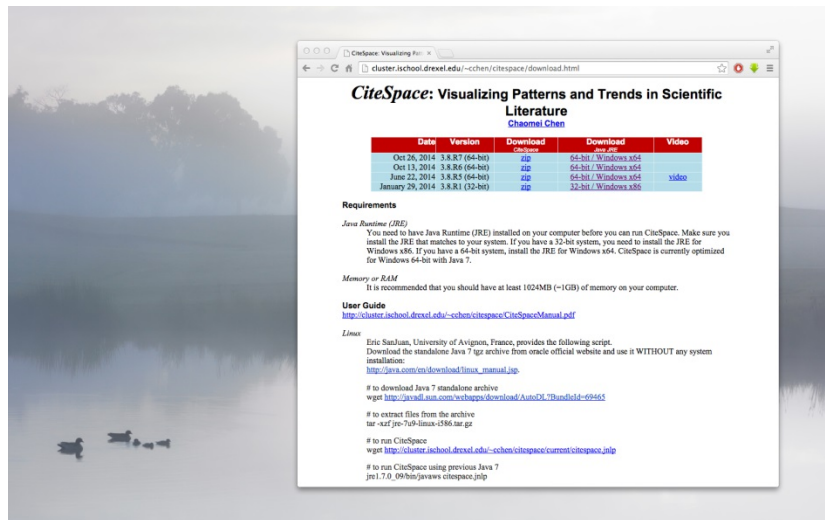


Figure 17. On a Mac, go to the CiteSpace home page in a browser such as Chrome and download the latest 64-bit version.

Once the download is completed, follow the option “Show in Finder.” It will take you to a list of files downloaded to your Mac. The most recent file should be the zip file for CiteSpace.

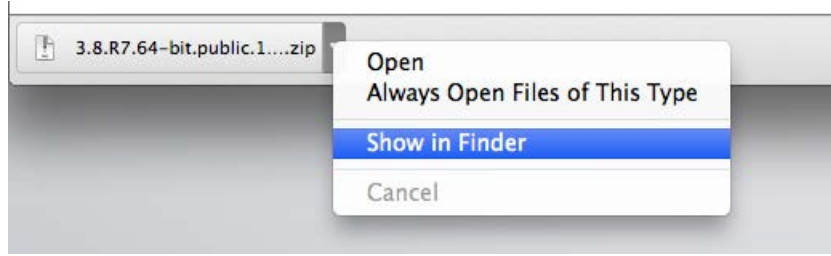


Figure 18. Choose “Show in Finder.”

Name	Date Modified	Size
 3.8.R7.64-bit.public.10.26.2014.zip	Today, 8:27 PM	25.4 MB

Figure 19. The downloaded zip file is shown in your Finder.

Double-click on the zip file to unzip the file to a folder in the current folder.

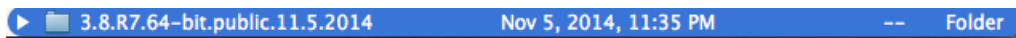


Figure 20. The zip file is unzipped to a new folder on the list.

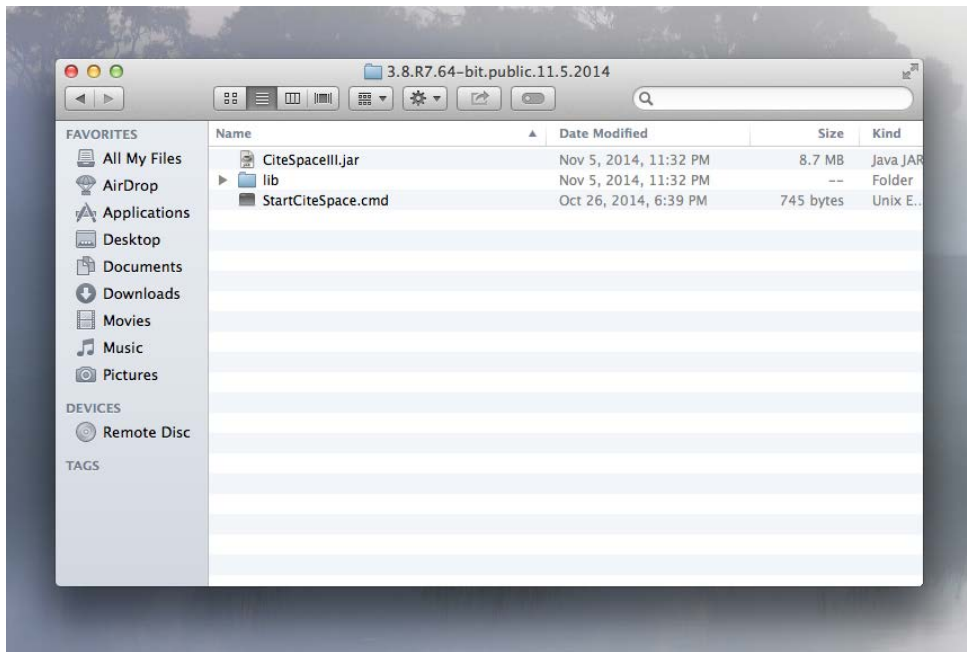
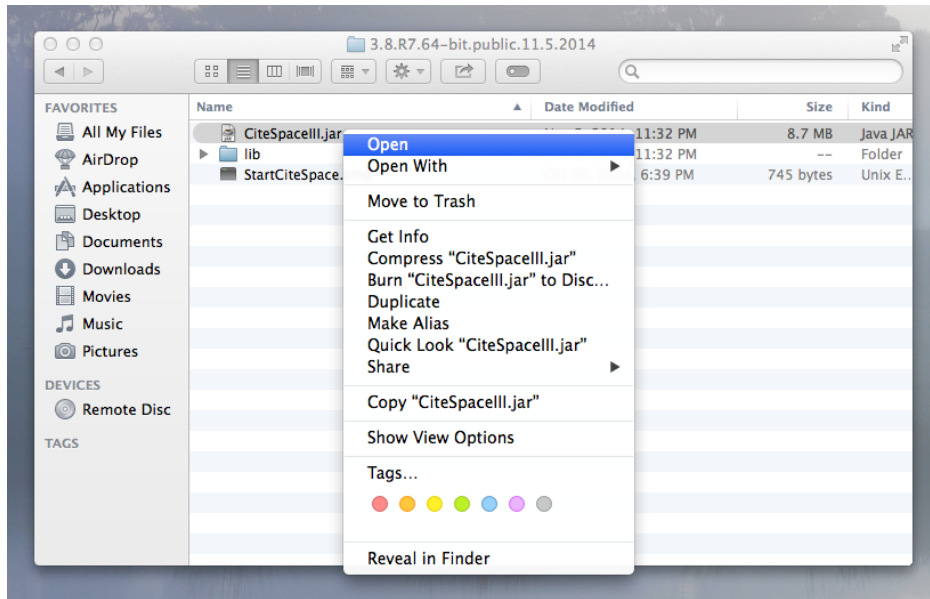


Figure 21. The new folder contains CiteSpaceII.jar and a lib folder.

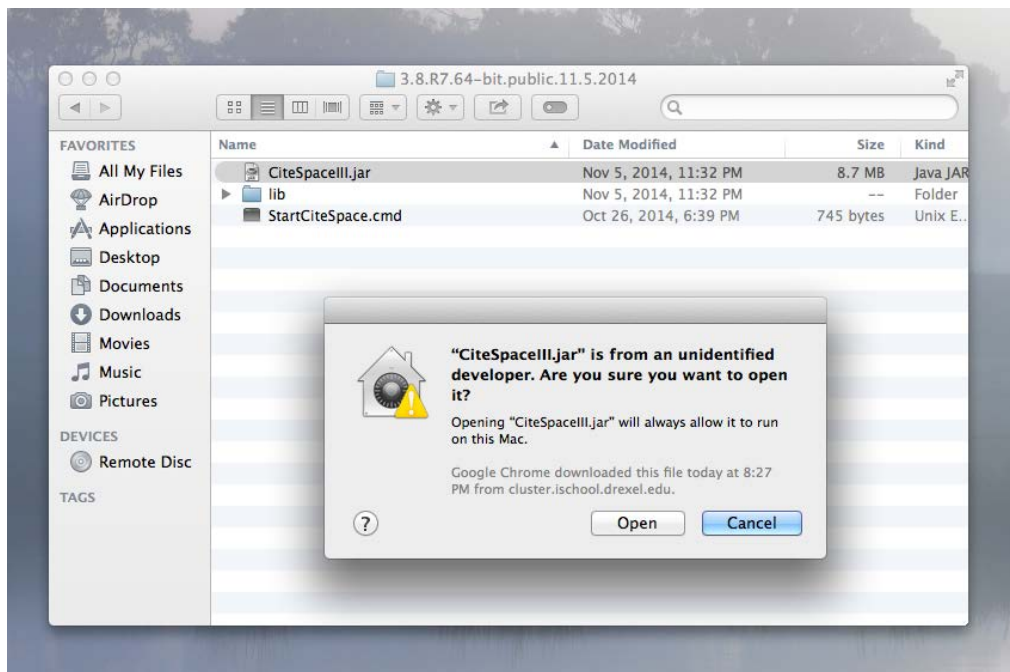
The simplest way to get started with CiteSpace is to open the CiteSpaceII.jar by clicking on it while holding the “Control” key on Mac. Select Open from the pop-up menu.





**Figure 22.** Click on the CiteSpaceII.jar while holding the “Control” key and select “Open.”

Due to the Java security settings, you will see a dialog box with two options for Open or Cancel. Choose Open to proceed. It will not harm your computer.



**Figure 23.** Choose “Open” from the dialog box to proceed.

After you choose Open, CiteSpace is getting started on Mac. You will see its opening page as follows. Choose “Agree” to continue.

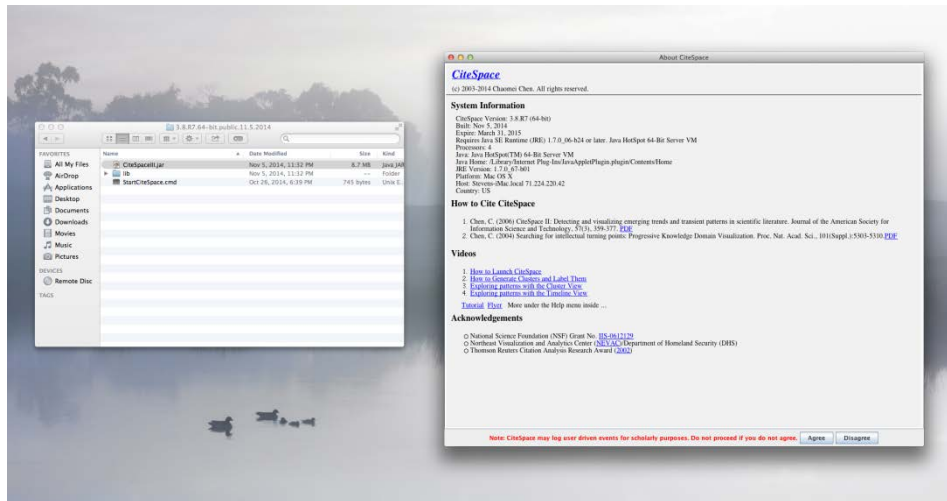


Figure 24. CiteSpace is now started on Mac.

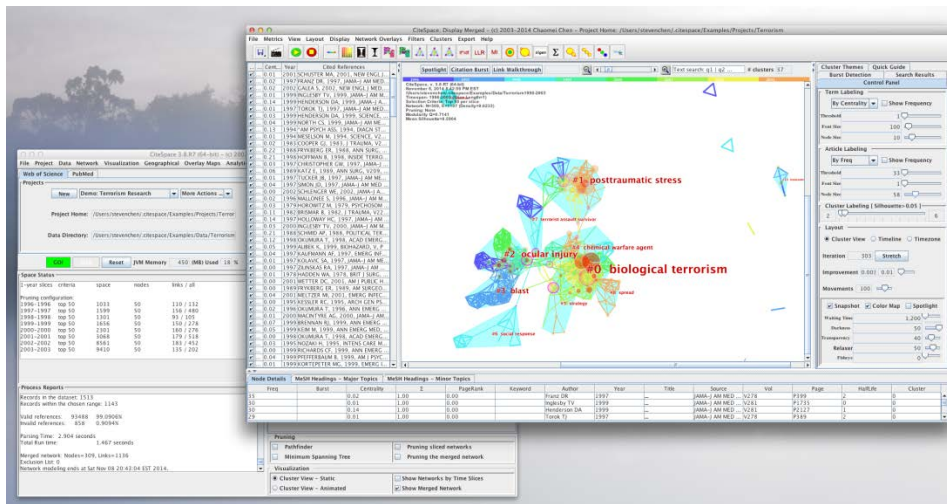


Figure 25. Screenshots of running the Demo project of CiteSpace on Mac.

It is a good idea to get familiar with the basic functions of CiteSpace by going through the Demo project on terrorism, which is included in the zip file.

If you want to configure various Java Virtual Machine parameters in more detail than what is shown in the above example, you may generate a bash file for your Mac as follows.

The Mac equivalent of the StartCiteSpace.cmd would be a bash file, which should have a file extension of .sh and should be executable. Let's name the file as StartCiteSpace.sh to be consistent.

1. The content of the StartCiteSpace.sh file should have the following two lines:

```
#!/bin/bash
java -Xms1g -Xmx4g -Xss5m -jar CiteSpaceIII.jar
```

2. The following instruction turns the StartCiteSpace.sh file to an executable file:

```
chmod +x StartCiteSpace.sh
```

3. To invoke the executable file, simply type its name or double click on it.



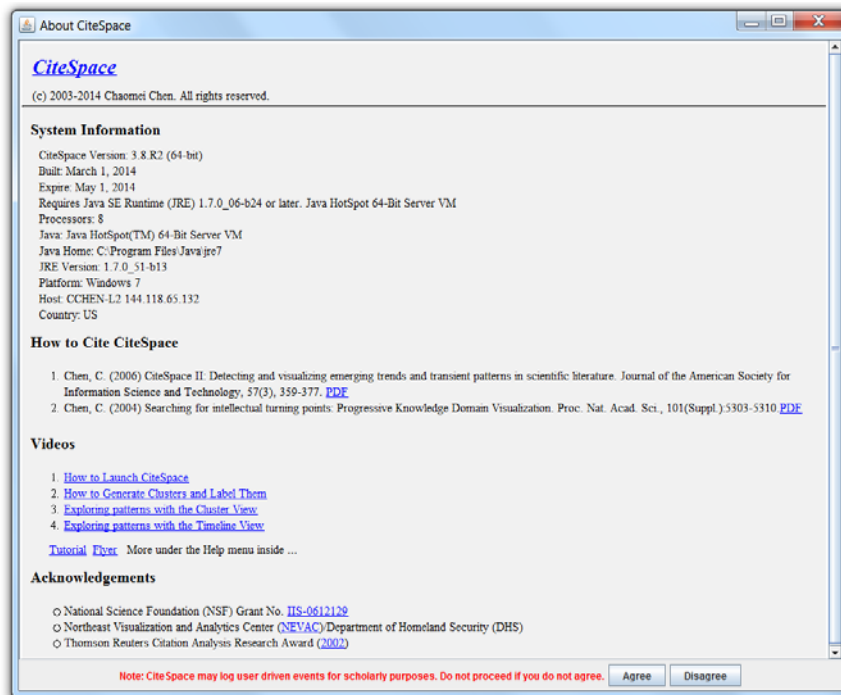


Figure 27. The “About CiteSpace” window. To proceed, click on the Agree button.

Next, you will see the main user interface of CiteSpace.

The user interface is divided into left and right halves. The left-hand side contains controls of projects (i.e. input datasets) and progress report windows. The right-hand side contains several panels for configuring the process with various parameters.

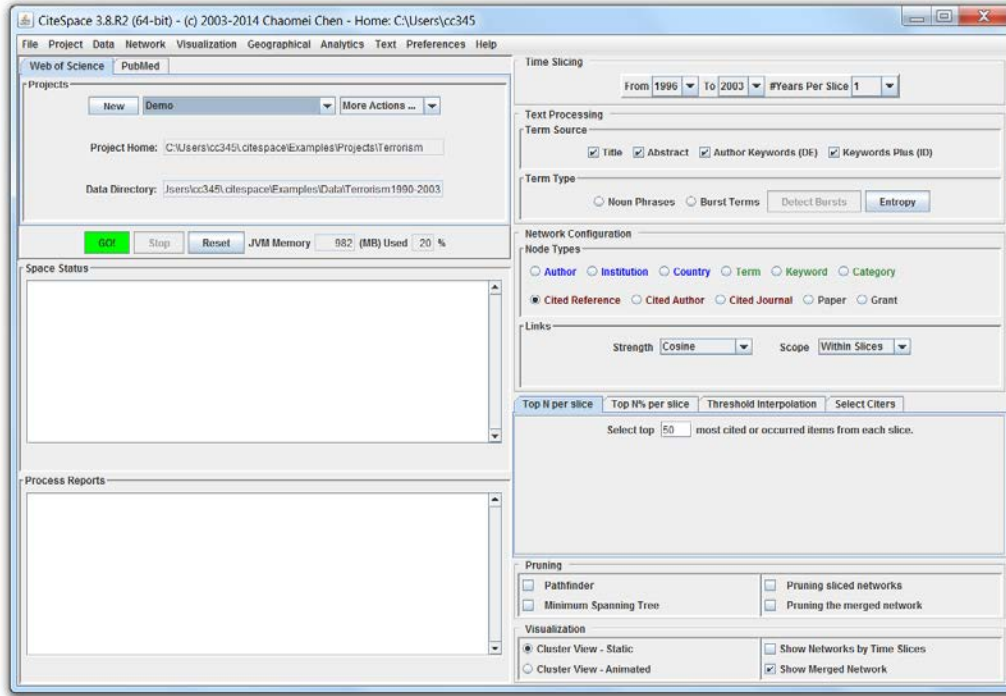
In a nutshell, the process in CiteSpace takes an input dataset specified in the current project, constructs network models of bibliographic entities, and visualizes the networks for interactive exploration for trends and patterns identified from the dataset.

The demo project contains a dataset on publications about terrorism research. These bibliographic records were retrieved from the Web of Science. See later sections on tips for how to construct your own dataset.

### 5.1.1 The Demo Project

We will start the process and explain how CiteSpace is designed to help you answer some of the key questions about a knowledge domain, i.e. a field of study, a research area, or a set of publications defined by the user.

Press the green GO! button to start the process.



**Figure 28. The main user interface of CiteSpace.**

CiteSpace will read the data files in the current project (Demo) and report its progress in the two windows on the left-hand side of the user interface. When the modeling process is completed, you have three options to choose: Visualize, Save As GraphML, or Cancel.

**Visualize:**

This option will take you to the visualization window for further interactive exploration.

**Save As GraphML:**

This option will save the constructed network in a file in a common graph format. No visualization.

**Cancel:**

This option will not generate any interactive visualization nor save any files. It allows you to reconfigure the process and re-run the process.

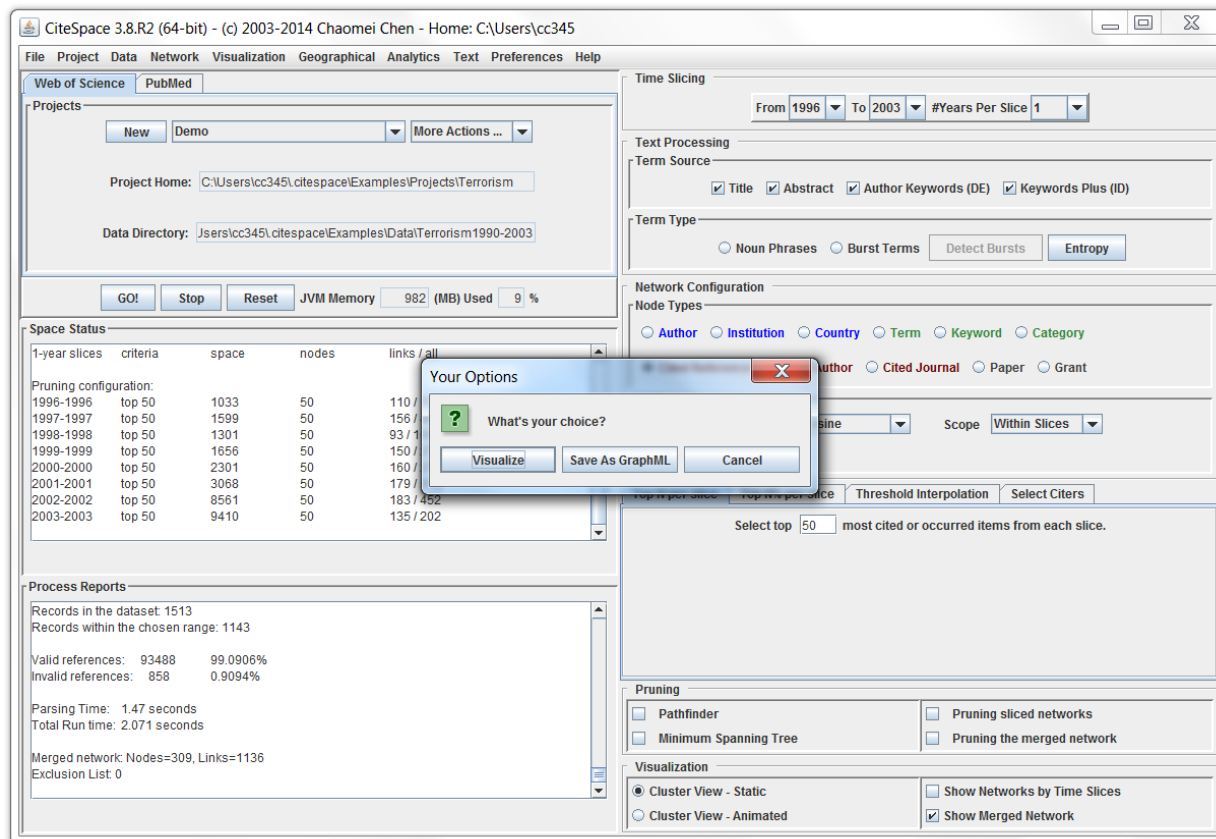


Figure 29. CiteSpace is ready to visualize the constructed network.

If you click on the Visualize button, a new window will pop up. This is the Visualization Window. Initially you will see some movements on your screen with a black background. Once the movements are settled, the background color turns to white.

Let's focus on what the initial visualization tells us and then explore what else we can find by using additional functions.

First, CiteSpace visualizes a merged network based on several networks corresponding to snapshots of consecutive years. In the Demo project example, the overall time span is from 1996 through 2003. The merged network characterizes the development of the field over time, showing the most important footprints of the related research activities. Each dot represents a node in the network. In the Demo case, the nodes are cited references. CiteSpace can generate networks of other types of entities. Here let's focus on cited references only for now. Lines that connect nodes are co-citation links; again, CiteSpace can generate networks of other types of links. The colors of these lines are designed to show when a connection was made for the first time. Note that this is influenced by the scope and the depth of the given dataset.

The color encoding makes it easy for us to tell which part of the network is old and which is new.

If you see that some references are shown with labels, then you will know that these references are highly cited, suggesting that they are probably landmark papers in the field. A list on the left side of the window shows all the nodes appeared in the visualization. The list can be sorted by the frequency of citations, Betweenness centrality, or by year or references as text. You can also choose to show or hide a node on the list.

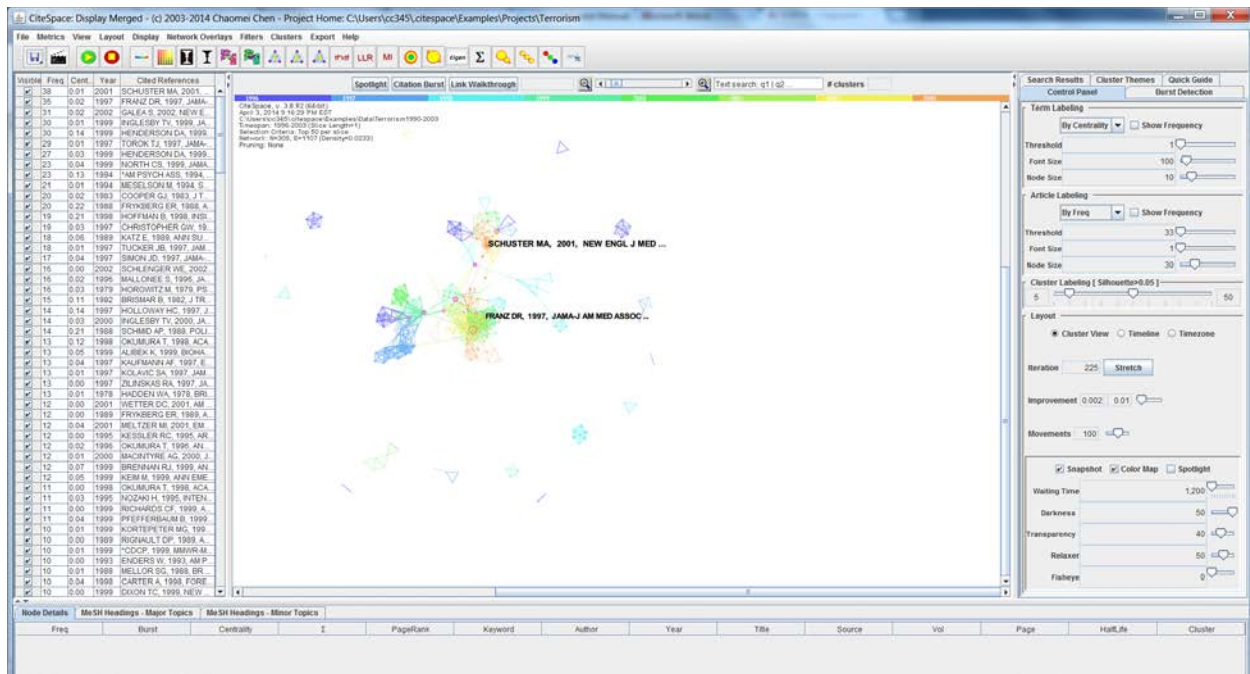


Figure 30. The Visualization window.

A control panel is shown on the right-hand side of the Visualization Window. You can change how node labels are displayed by a combination of a few threshold values through sliders. You can also change the size of a node by sliding the node size slider.

To answer the typical questions we asked before, let's use several functions in CiteSpace to obtain more specific information through clustering, labeling, and exploring.

### 5.1.2 Clustering

Although we can probably eyeball the visualized network and identify some prominent groupings, CiteSpace provides more precise ways to identify groupings, or clusters, using the clustering function.

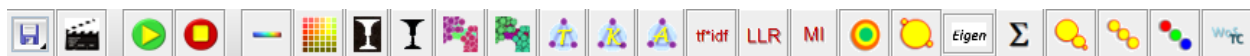



Figure 31. Most frequently used functions for visual exploration in CiteSpace.

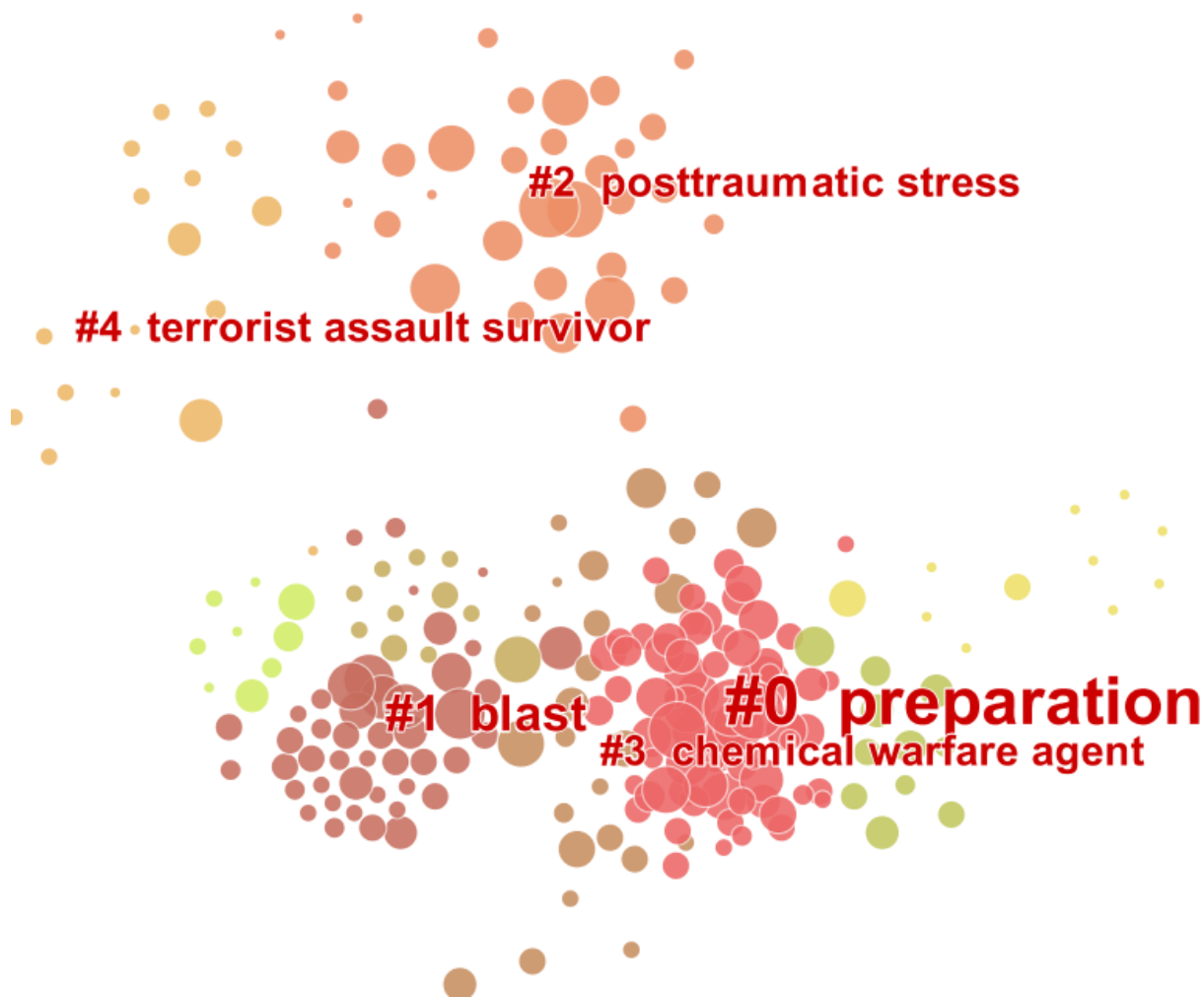
To start the clustering function, simply click on this icon .

How do I know whether the clustering process is completed? You will see #clusters on the upper right corner of the canvas. In the Demo example, a total of 37 clusters of co-cited references are identified. Each cluster corresponds to an underlying theme, a topic, or a line of research.

The signature of the network is shown on the upper left corner of the display. In particular, the modularity  $Q$  and the mean silhouette scores are two important metrics that tell us about the overall structural properties of the network. For example, the modularity  $Q$  of 0.7141 is relatively high, which means that the network is reasonably divided into loosely coupled clusters. The mean silhouette score of 0.5904 suggests that the homogeneity of these clusters on average is not very high, but not very low either.



**Figure 32.** The clustering process is completed. 37 clusters are identified (#clusters shown in the upper right corner). Modularity and silhouette scores are shown in the signature of the network on the left.



**Figure 33.** Members of different clusters are shown in different colors.

You can inspect various measures of each cluster in a summary table of all the clusters using: Clusters ► 4. Summarization of Clusters. The Silhouette column shows the homogeneity of a cluster. The higher the silhouette score, the more consistent of the cluster members are, provided the clusters in comparison have similar sizes. If the cluster size is small, then a high homogeneity does not mean much. For example, cluster #9 has 7 members and a silhouette of 1.00, this is most likely due to the possibility that all 7 references are the citation references of the same underlying author. In other words, cluster #9 may reflect the citing behavior or preferences of a single paper, thus it is less representative.

The average year of publication of a cluster indicates whether it is formed by generally recent papers or old papers. This is a simple and useful indicator.



Select	Cluste...	Size	Silhou...	mean(...)	Top Terms (tf*idf weighting)	Top Terms (log-likelihood ratio, p-level)	Terms (mutual information)
<input type="checkbox"/>	0	65	0.651	1996	(16.48) biological terrorism; (15.97) ...	biological terrorism (66.82, 1.0E-4); s...	nuclear terrorism
<input type="checkbox"/>	1	37	0.92	1995	(18.54) posttraumatic stress; (17.1) tr...	september (116.08, 1.0E-4); terrorist ...	history
<input type="checkbox"/>	2	36	0.9	1987	(15.8) ocular injury; (15.14) eye injury...	oklahoma city bombing (94.73, 1.0E-...	terror defense
<input type="checkbox"/>	3	26	0.818	1982	(14.97) blast; (14.65) blast over-pres...	blast (79.4, 1.0E-4); blast over-pres...	blast injury
<input type="checkbox"/>	4	24	0.815	1995	(11.96) chemical warfare agent; (11.9...	emergency (48.82, 1.0E-4); chemical ...	nuclear terrorism
<input type="checkbox"/>	5	14	0.886	1997	(10.94) strategy; (9.62) architecture; (...)	government (18.67, 1.0E-4); architect...	history
<input type="checkbox"/>	6	13	0.983	1990	(11.96) social response; (11.96) bas...	social response (33.9, 1.0E-4); basq...	terror
<input type="checkbox"/>	7	12	0.901	1989	(12.8) terrorist assault survivor; (12.8)...	terrorist assault survivor (37.89, 1.0E-...	unabomber
<input type="checkbox"/>	8	11	0.969	1999	(15.14) spread; (14.6) smallpox; (12.8)...	smallpox (106.47, 1.0E-4); spread (3...	terror defense
<input type="checkbox"/>	9	7	1	1987	(12.8) abolition; (12.8) nuclear war; (1...	destruction (53.05, 1.0E-4); medicine ...	medical care
<input type="checkbox"/>	10	7	1	1988	(11.96) indigenous guatemalan refug...	indigenous guatemalan refugee child...	analysis
<input type="checkbox"/>	11	7	1	1991	(9.62) repression; (9.62) dynamic mo...	repression (24.27, 1.0E-4); dynamic ...	21st century
<input type="checkbox"/>	12	6	1	1988	(12.8) american terrorist state; (12.8) ...	american terrorist state (50.88, 1.0E-...	effect
<input type="checkbox"/>	13	5	1	1990	(6.53) transnational terrorism; (4.21) L...	transnational terrorism (21.74, 1.0E-4...	transnational terrorism

Figure 34. A summary table of clusters.

### 5.1.3 Generate Cluster Labels

To characterize the nature of an identified cluster, CiteSpace can extract noun phrases from the titles (T in the following icon), keyword lists (K), or abstracts (A) of articles that cited the particular cluster.

Let's ask CiteSpace to choose noun phrases from titles (i.e. select the T icon). This process may take a while as CiteSpace needs to compute several selection metrics. Once the process is finished, the chosen labels will be displayed. By default, labels based on one of the three selection algorithms will be shown, namely, tf\*idf. Our study has found that LLR usually gives the best result in terms of the uniqueness and coverage.



Figure 35. Icons for performing Clustering and Labeling functions.

Cluster labels are displayed once the process is completed. The clusters are numbered in the descending order of the cluster size, starting from the largest cluster #0, the second largest #1, and so on.

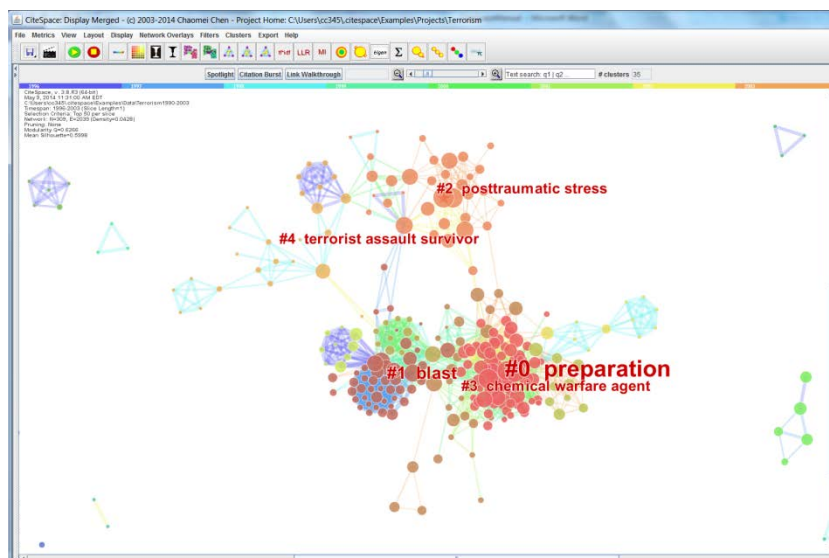


Figure 36. Cluster labels are generated and displayed.

To make it easier to see which clusters are the largest, you can choose to change the font size of the labels from the uniformed to proportional:

Display ► Label Font Size ► Cluster: Uniformed/Proportional

This is a toggle function. That means there are two states. Your selection will switch back and forth between the two states, i.e. either using a uniformed font size or proportional.

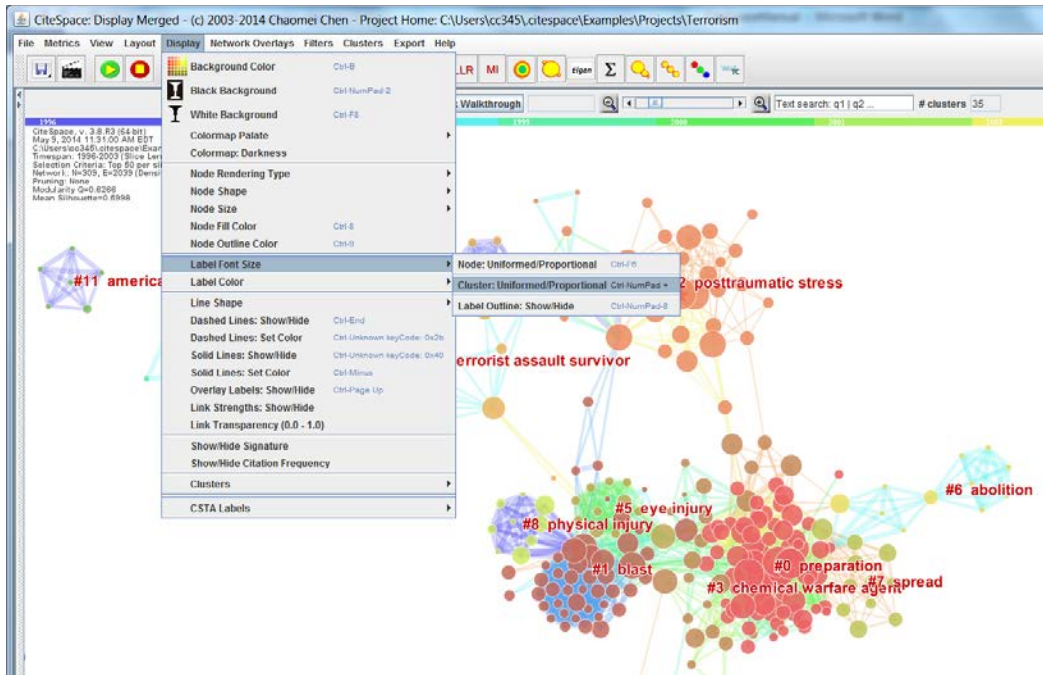


Figure 37. Set the cluster labels' font size proportional to their size.

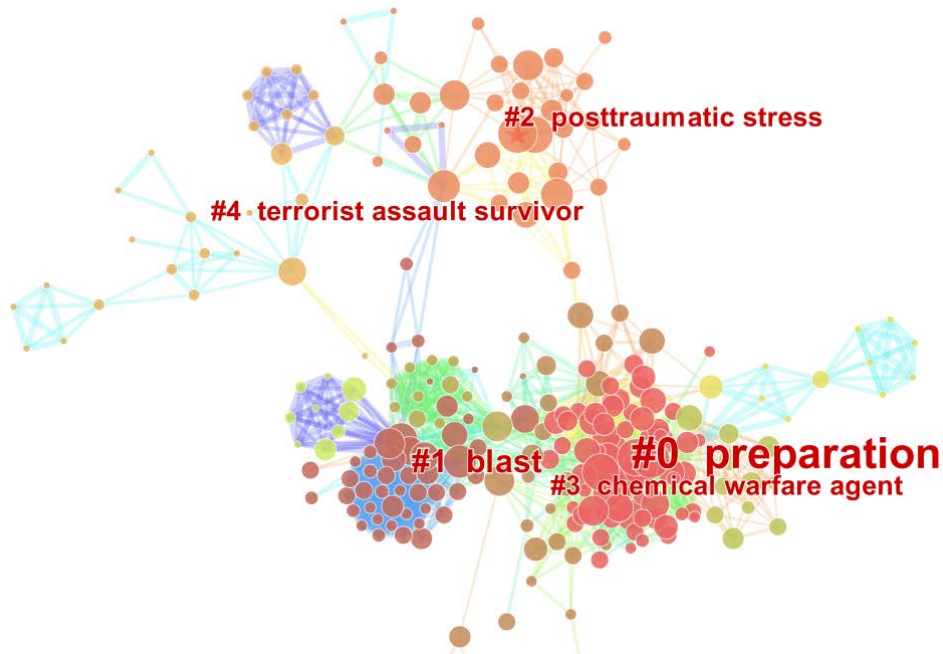


Figure 38. Cluster labels' font sizes are proportional to the size of a cluster. The largest cluster is #0 on biological terrorism.

### 5.1.4 Where are the major areas of research based on the input dataset?

This is one of the primary questions that CiteSpace is designed to answer. To answer this question, we will focus on the big picture of the collection of publications represented by your dataset. Let's make a few adjustments with the sliders in the control panel on the right so that the information of our interest will be shown clearly and information that is less relevant to this question right now will be temporarily hidden from the view.

#### 1. Node Size

At this level, we don't really need to see the size of a node, although it provides rich information about the history of a node. Use the slider under **Article Labeling** ► **Node Size** ► [Slide to 0] (marked by the pointer #1 in the following figure).

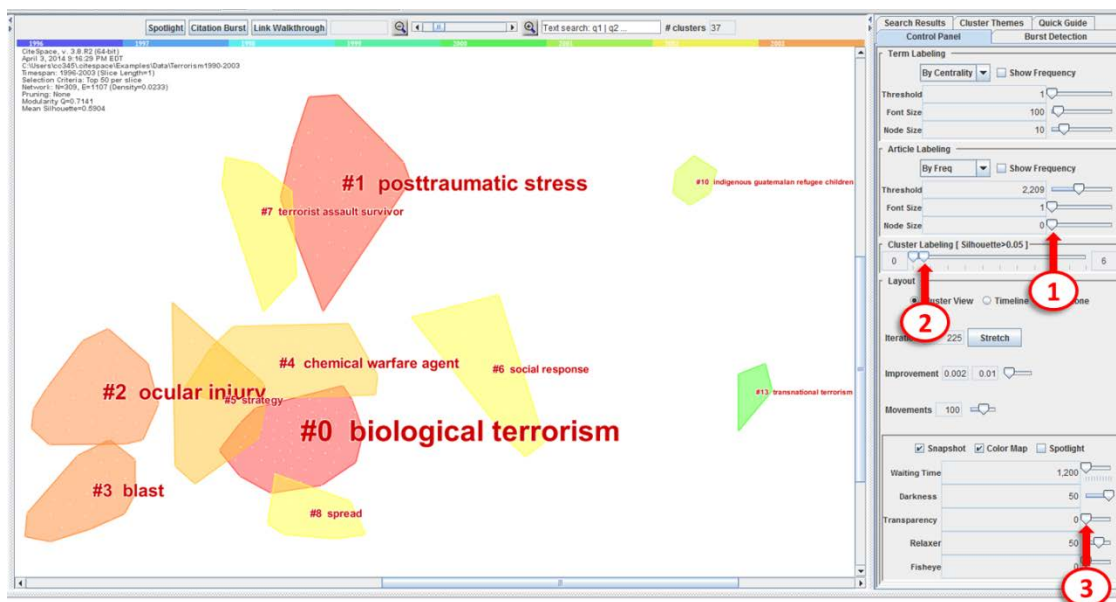
#### 2. Cluster Label Size

The font size of the cluster labels are controlled by a slider with two controls: one control the threshold for showing or hiding a label based on the size of the cluster (i.e. to make sure large-enough clusters are always labeled), and the other control the font size of the cluster labels (marked by the pointer #2 in the screenshot).

#### 3. Transparency of Links

Detailed links would be useful later, but we can ignore them for now using the transparency slider to set all the links' transparency to the lowest level, i.e. invisible. In hindsight, a more accurate term would be completely transparent.

After making these minor adjustments, it will be straightforward to answer the question: Where are the major areas of research? Evidently, the largest area (cluster #0 with the largest number of member references) is biological terrorism. The second largest is posttraumatic stress (cluster #1), i.e. PTSD. The third one is ocular injury (cluster #2). The fourth one is blast (cluster #3). And there are a few smaller clusters. So now we have a general idea what constituted terrorism research during the period of 1996 and 2003. You can repeat the process on a current dataset to get an up-to-date big picture.



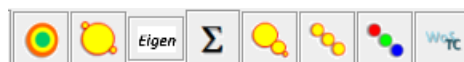
**Figure 39. Adjust the appearance of the visualization with a few sliders. Pointers: 1) Node size control slider, 2) cluster label size, and 3) transparency of links.**

### 5.1.5 How are these major areas connected?

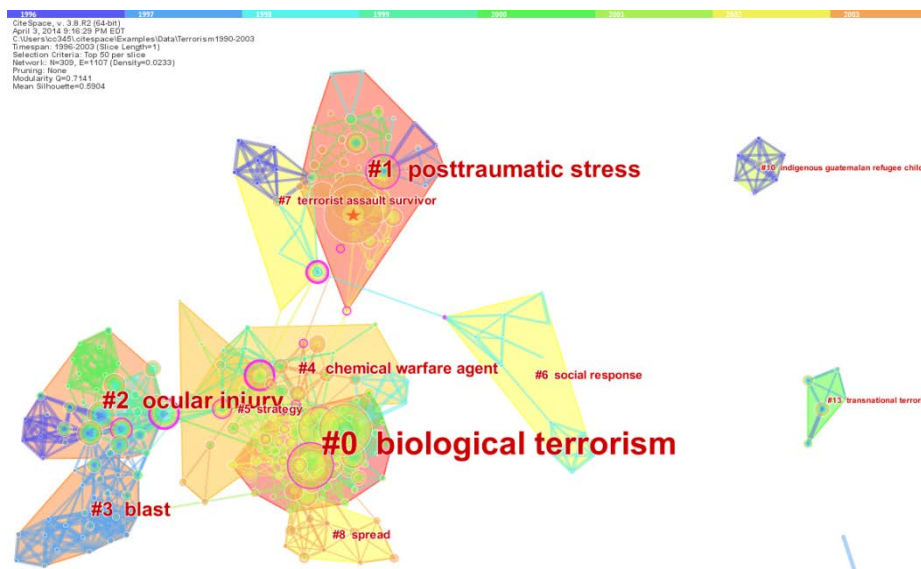
To answer this question, we need to bring back the lines connecting nodes. Adjust the transparency slider to make the lines visible.

A useful indicator of how different clusters are connected is a type of nodes that have high betweenness centrality scores. In CiteSpace, betweenness centrality scores are normalized to the unit interval of [0, 1]. A node of high betweenness centrality is usually one that connects two or more large groups of nodes with the node itself in-between, hence the term betweenness. CiteSpace highlights nodes with high betweenness centrality with purple trims. The thickness of a purple betweenness centrality trim indicates how strong its betweenness centrality is. The thicker the stronger. Occasionally, a node with high betweenness centrality may appear at the center of a network component, but our interest is in the nodes that are truly in between.

To make see the purple rings, switch the node rendering mode to tree rings, which is the first icon shown in the following figure, i.e. concentric citation rings represent how many citations were made to the node in corresponding years. Remember that colors represent when citations were actually made.



**Figure 40. Icons of node rendering controls.**



**Figure 41. Nodes with purple rings are important in connecting different clusters.**

### 5.1.6 Where are the most active areas?

#### 5.1.6.1 Burst Detection

Citation burst is an indicator of a most active area of research. Citation burst is a detection of a burst event, which can last for multiple years as well as a single year. A citation burst provides evidence that a particular publication is associated with a surge of citations. In other words, the publication evidently has attracted an extraordinary degree of attention from its scientific

community. Furthermore, if a cluster contains numerous nodes with strong citation bursts, then the cluster as a whole captures an active area of research, or an emerging trend.

The burst detection in CiteSpace is based on Kleinberg’s algorithm (Kleinberg, 2002).

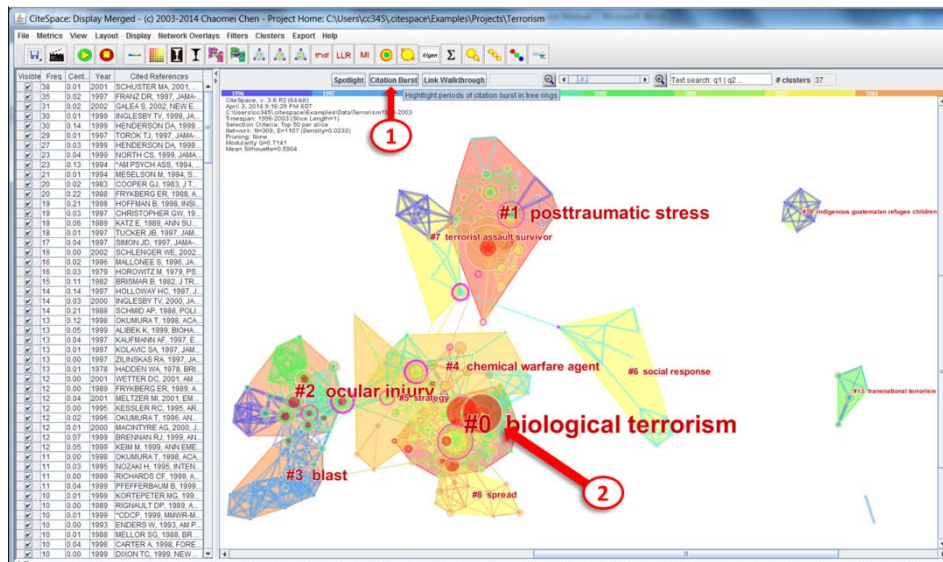


Figure 42. Citation bursts are indicators of most active areas.

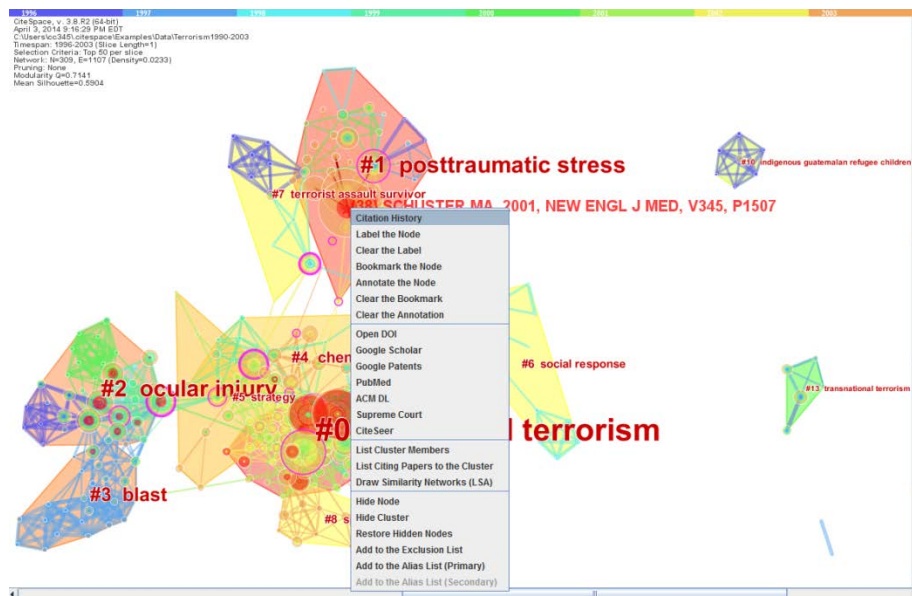
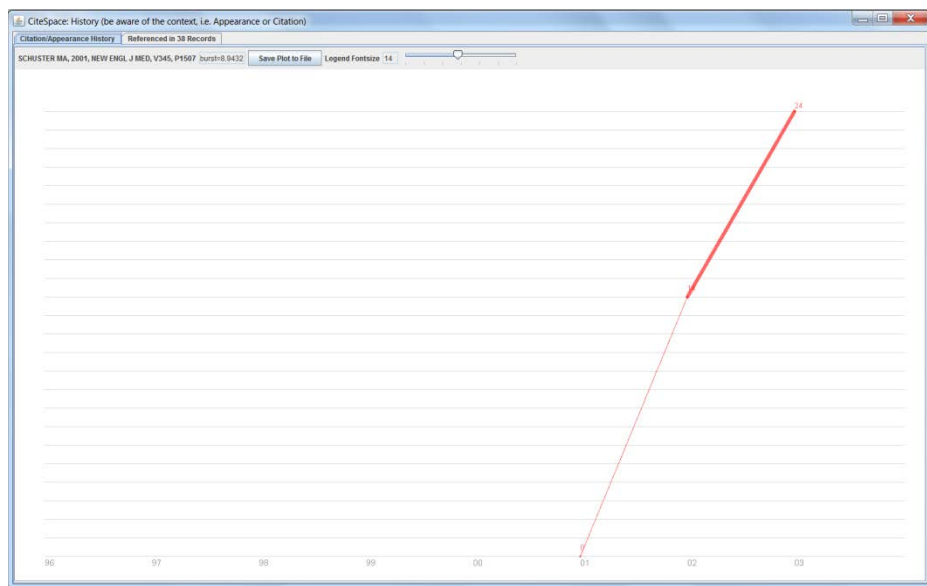


Figure 43. Right click on the node of interest and choose the Citation History of the node.



**Figure 44.** This is the citation history of an article that has a citation burst.

Using **View ► Citation Burst History** can generate a summary list of articles that are associated with citation bursts. This visualization shows which references have the strongest citation bursts and which periods of time the strongest bursts took place. For example, from the list, we can tell that Schuster et al. (2001) has the strongest bursts among articles published since terrorist attacks in 2001. It is also interesting to note that North et al. (1999) has the second strongest citation burst in the period of 2002 and 2003.

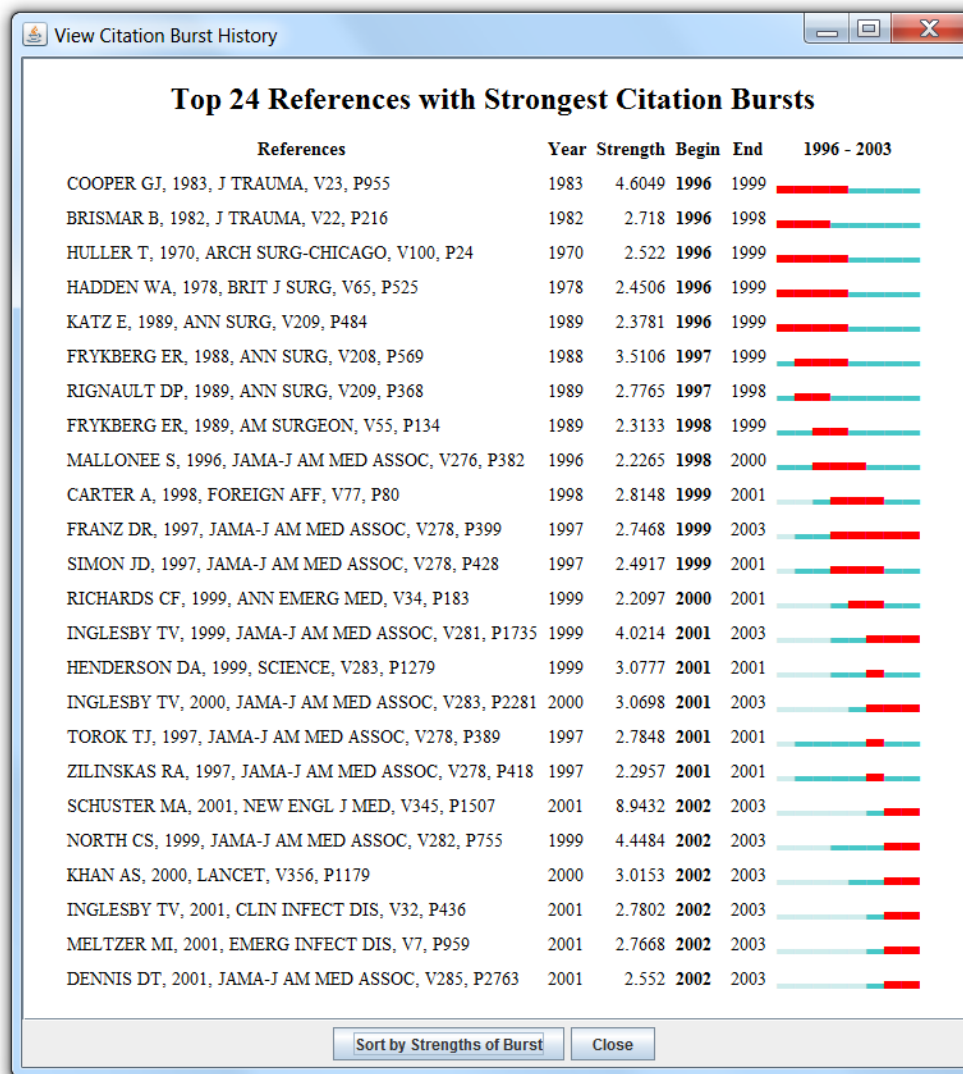


Figure 45. A summary list of references with citation bursts.

Burst detection and visualization can be applied to other types of nodes. For the node type of author, it will show you those authors who have rapidly increased the number of publications. Similarly, institutions will identify universities that are particularly active in the relevant research areas. For keywords, it will show you fast growing topics.

The general procedure is the same for different types of nodes. Here we illustrate the procedure with an example of detecting the burstness of keywords in publications of Drexel University between 2000 and 2014.

1. Select the node type: **Keyword**
2. Generate a network as usual: **2000-2014; Slice length: 1; Top N=100; GO**
  - o (N=392, E=3033)
3. Run the burst detection function: **Citation Burst**
4. Visualize the entities, i.e. nodes, that have bursts: **View > Citation Burst History**

### Top 30 Keywords with Strongest Citation Bursts

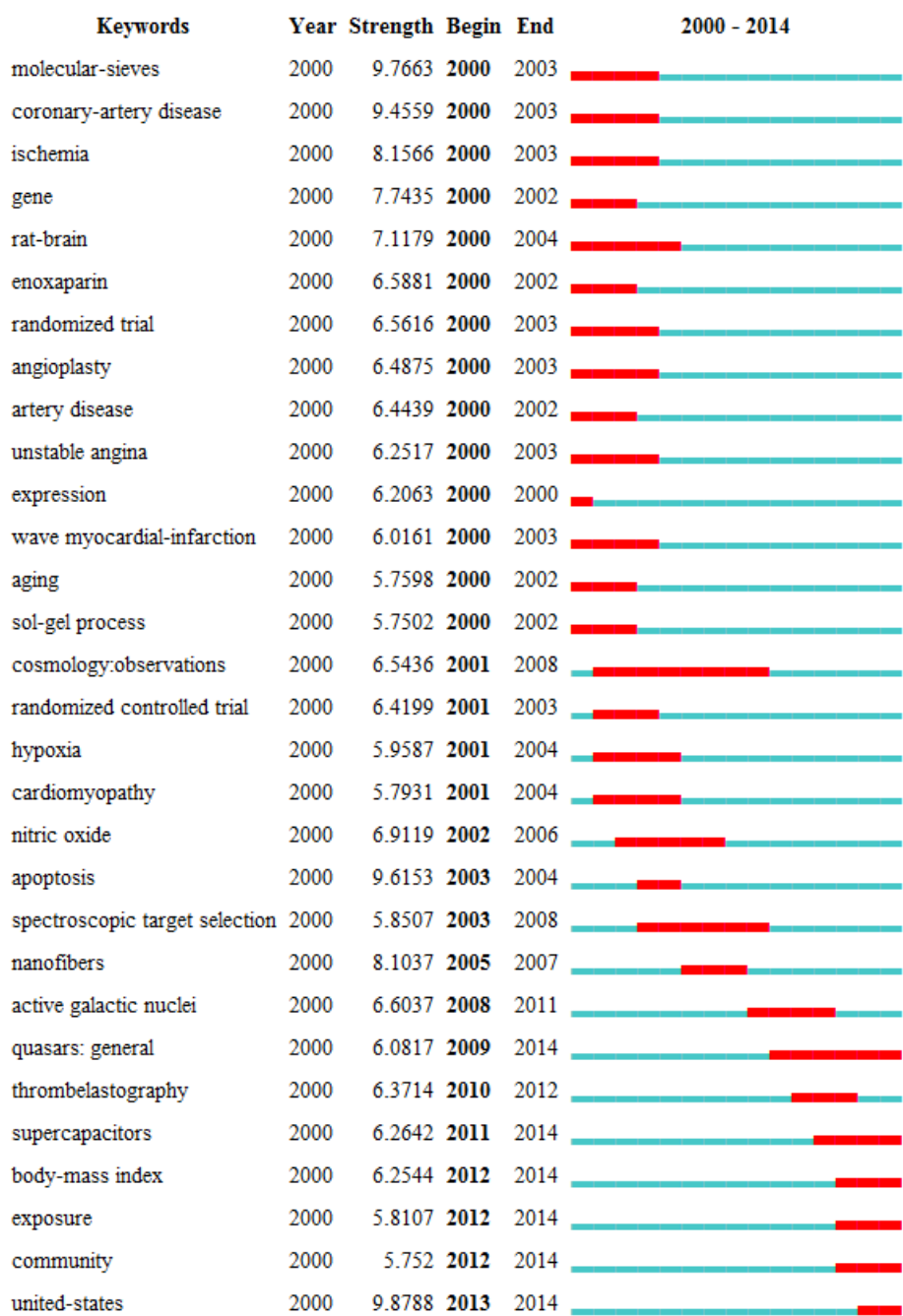


Figure 46. A visualization of the history of the burstness of keywords in publications of Drexel University (2000-2014). For example, cosmology: observations in Astronomy has the longest period of burst from 2001 till 2008.

#### 5.1.6.2 Burst Detection – Additional Controls

If the number of burst items is too many or too few, you can adjust several parameters for the burst detection algorithm.



I will illustrate the steps for detecting bursts of authors, i.e. authors who have published at a very fast rate, with the terrorism dataset that comes with the CiteSpace package. The procedure is the same for other node types.

First, select Author in the Node Types panel and unselect other node types.

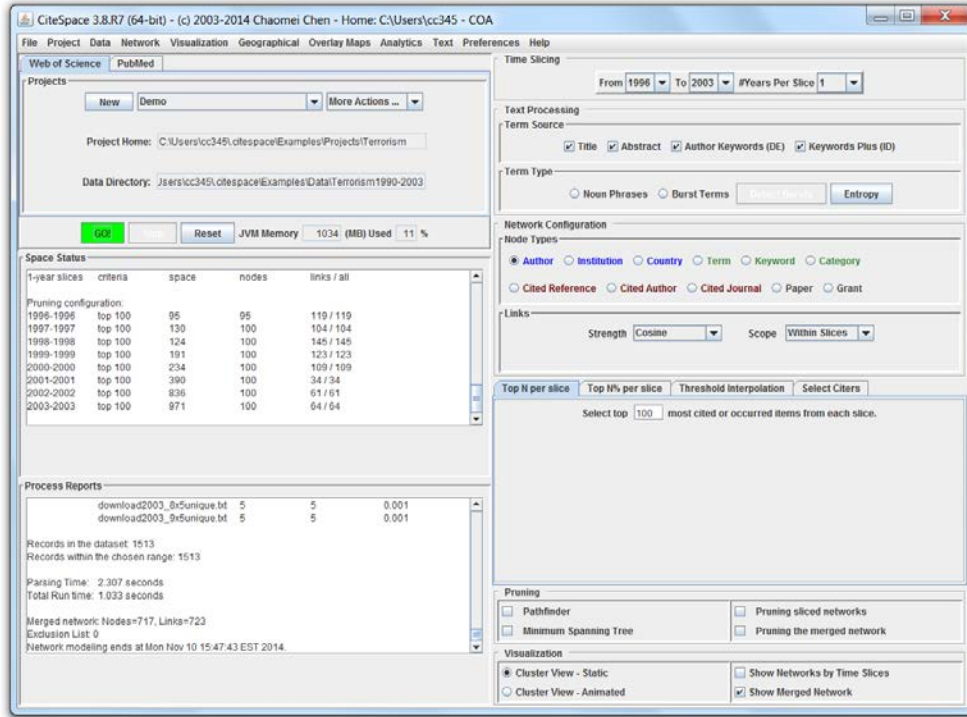


Figure 47. Select Author as the Node Type and unselect other types.

Wait until the Display window shows up. The layout process may continue to run for a while. Since our focus is on the burst detection, you can stop the layout process anytime you like by clicking on the Stop button (which is the one with a yellow square on a red background). Next, click on the Citation Burst button.

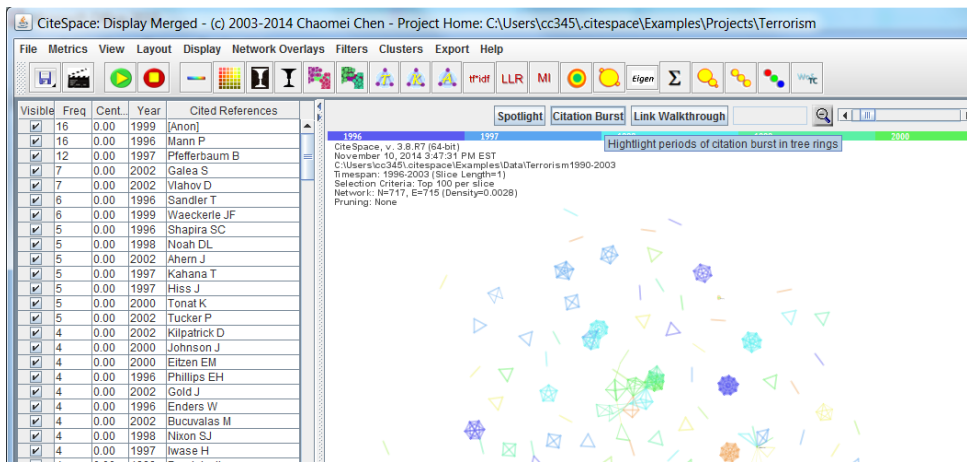


Figure 48. Once the layout process is completed, i.e. when you see the background turns to white, click on the Citation Burst button. You can also force the layout process to stop by click on the Stop button.



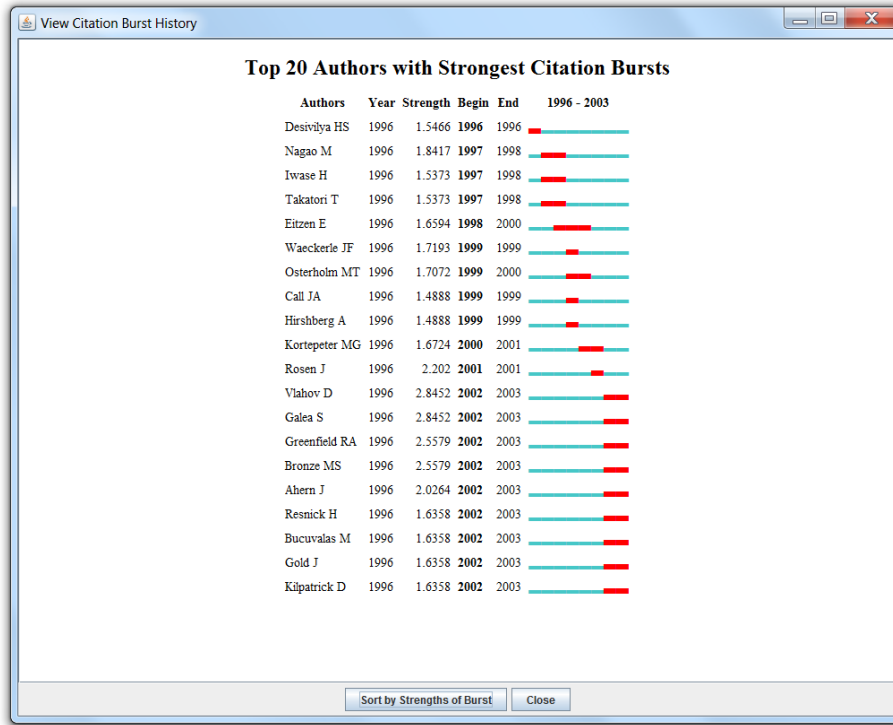


Figure 52. Top 20 Authors with strongest bursts, i.e. who were most active in published papers according to the dataset.

If the initial burst detection only identifies a small number of items, you may adjust the parameters provided in the Burst Detection panel to increase or decrease the total number of burst items. For example, the following combination of parameters increases the number of authors with burst patterns from 37 to 85.

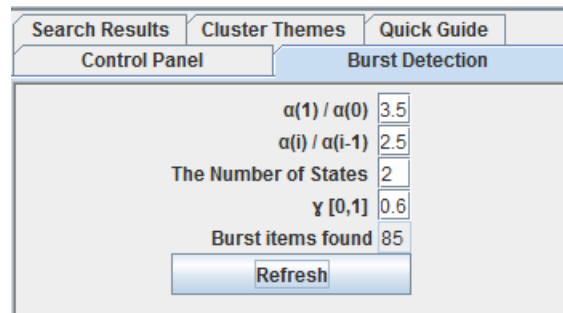


Figure 53. Four additional parameters on the Burst Detection panel can be adjusted to control the burst detection algorithm.

After clicking on the Refresh button, CiteSpace will re-calculate the burstness of all the items. Then you can use the steps described above to display the new results.

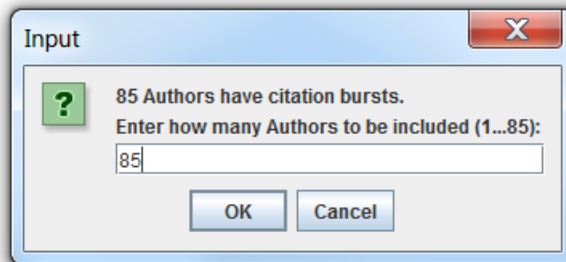


Figure 54. The re-calculated burst detection found 85 authors, a substantial increase from the 37 authors found by the default setting.

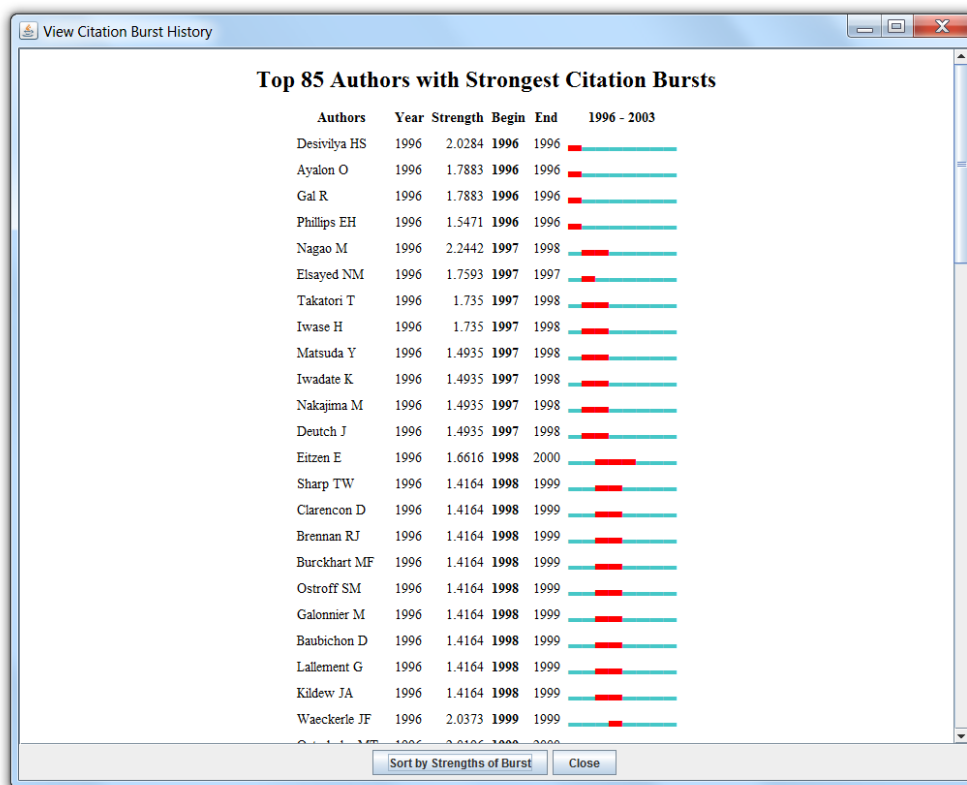


Figure 55. The list shows the new results from a re-configured burst detection process.

### 5.1.7 What is each major area about? Which/where are the key papers for a given area?

Cluster labels can tell us the context in which they are most cited because the label terms are extracted from citing articles' titles, keywords, or abstracts.

To explore these clusters in more depth, you should use the Cluster Explorer:

Clusters ► Cluster Explorer

The initial appearance of the Cluster Explorer shows four windows: 1) Clusters, 2) Citing Articles, 3) Cited References, and 4) Representative Sentences. Windows 2-3 are blank until you select a cluster in the Clusters window by checking the checkbox in front of each row of cluster information.



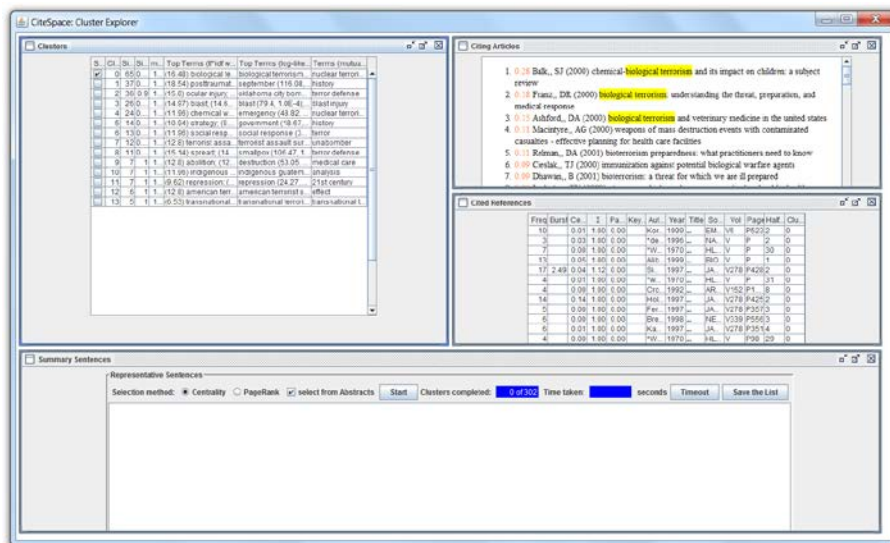


Figure 57. Cluster Explorer: Cluster #0 is selected in the checkbox.

In the Summary Sentences window, if you click on the Start button, CiteSpace will extract the most representative sentences from the abstracts of the citing articles to each cluster. A sentence is considered representative if it is either a sentence with a high degree centrality or a sentence with a high PageRank score.

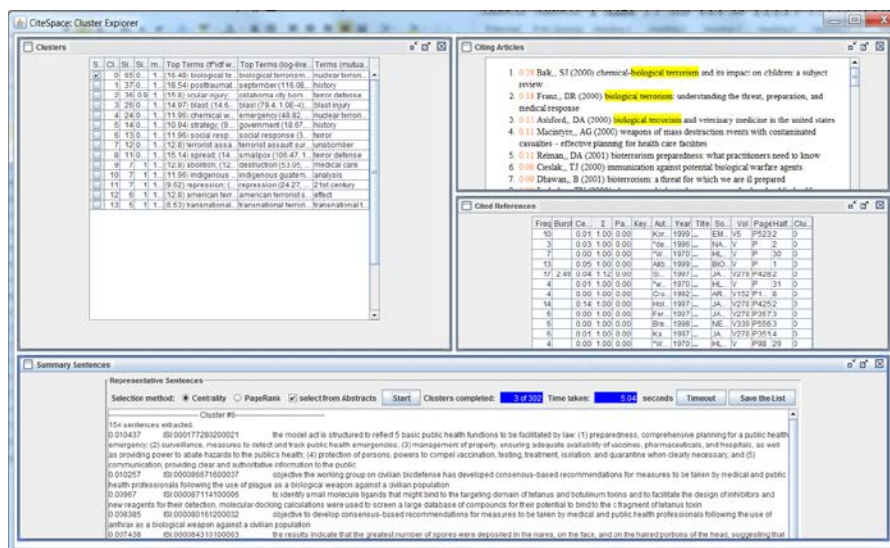


Figure 58. Representative sentences are displayed upon clicking on the Start button in the Summary Sentences window.

### 5.1.8 Timeline View

You can switch to a timeline view of the network by choosing the Timeline radio button in the Layout panel on the right (as pointed by the red arrow in the following figure). In a timeline view, each cluster is arranged on a horizontal timeline. The direction of time points to the right.

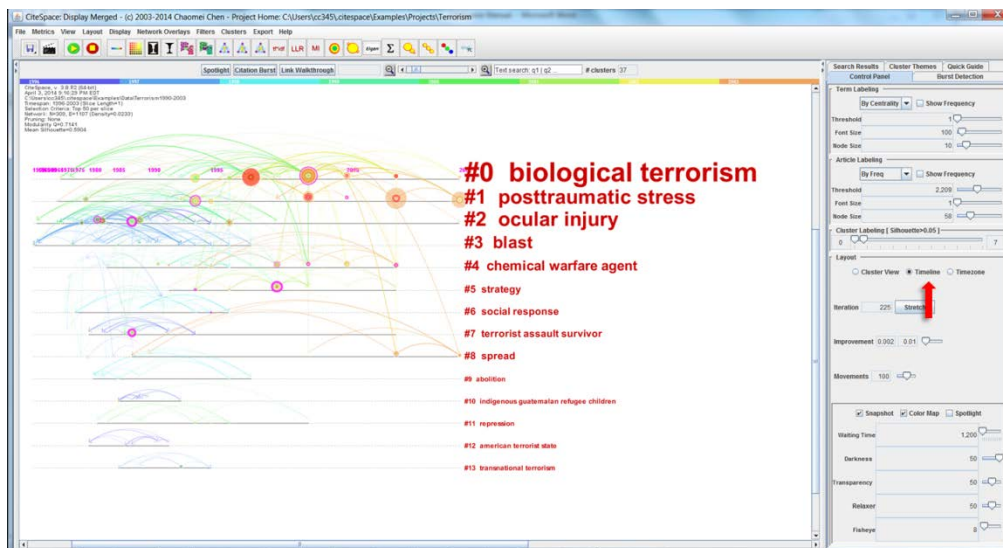


Figure 59. A timeline view of the network.

You have seen some of the basic moves. CiteSpace has many other features. We will introduce other features at more advanced levels.

## 5.2 Try it with a dataset of your own

### 5.2.1 Collecting Data

#### 5.2.1.1 How to construct my own data from the Web of Science

The primary source of data for CiteSpace is the Web of Science.

Most importantly, the dataset should include cited references in order to maximize the potential of CiteSpace.

The Web of Science has several ways to search for bibliographic records. The most basic one is called, of course, basic search, which includes topic, author, and several other searchable fields. The following example shows a topic search for “CiteSpace” between the timespan of 2004 and 2014.

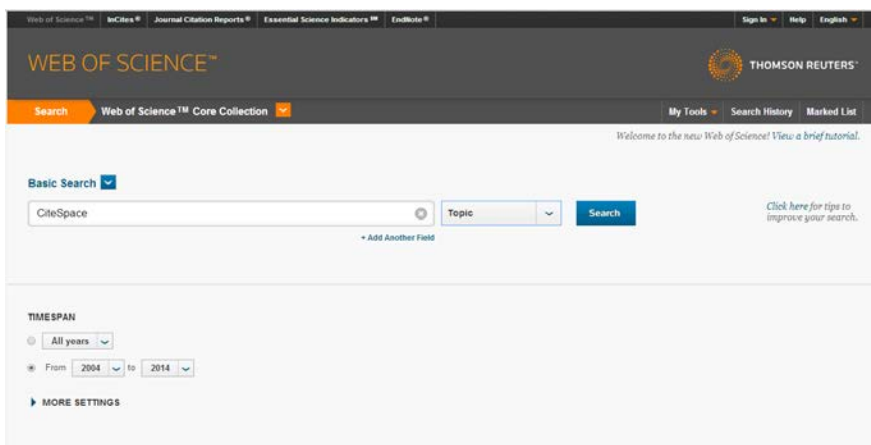


Figure 60. A topic search in the Web of Science.

The topic search found 16 results. The results are initially displayed in the chronological order of the publication date from the newest to the oldest. You can switch to a different order, for example, by the number of citations, from the highest to the lowest, so you can quickly narrow down to a small subset of the most highly cited records.

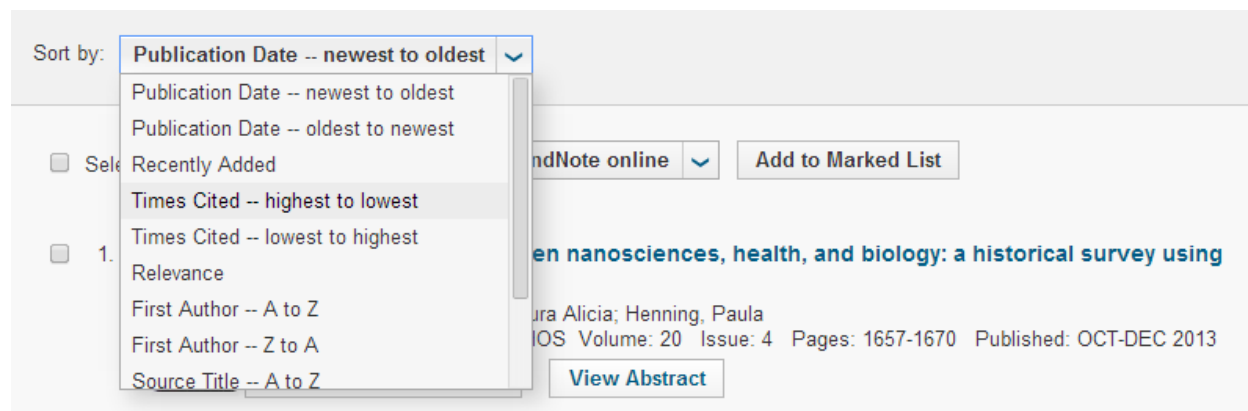


Figure 61. Sort the results by Times Cited – highest to lowest.

You will notice if the results are sorted by Times Cited – highest to lowest. The record with the highest times cited is the 2006 JASIST paper on CiteSpace II, with 185 citations. The topic search found 16 records. You can download these 16 records, however, that would be not representative. If you follow the Create Citation Report link, you will see you can expand the 16 records to about 220 records that cited the set of 16 records. We refer to this way to obtain more potentially relevant records as citation expansion. Since the only thing we know is that each record in the expanded set at least cited one of the original 16 records, it may turn out to be a less relevant record because of the diversity of how authors cite. Let's if we can do better than finding 220 records related by citation indexing.

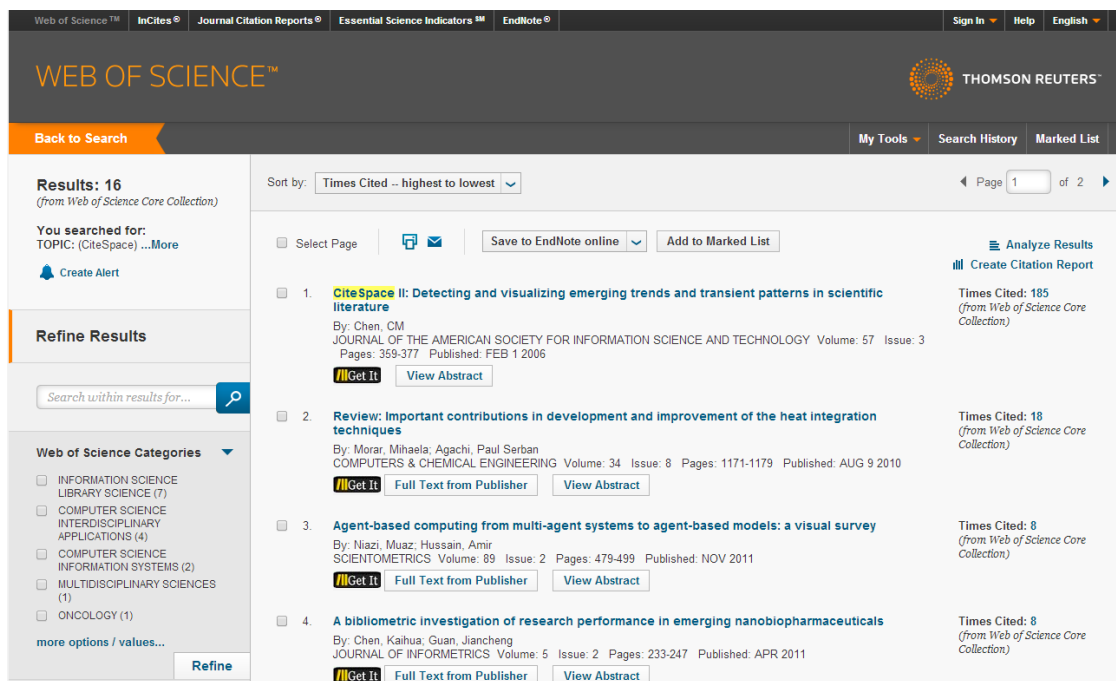


Figure 62. Results are now sorted by Times Cited from the highest to the lowest.



You may also notice that the 2004 PNAS and the 2010 JASIST paper on CiteSpace were NOT on the list, although they are certainly about CiteSpace and their citations would put them on the list too. Thus, this example shows that you should be careful when using the topic search along to construct your own dataset.

Under the Citation Network panel, the 104 Times Cited is a clickable link. If you click on it, it will bring you to the list of 104 records that cited the 2004 PNAS paper. The 2006 JASIST paper should be on the list. If we sort the list by Times Cited, then we will see the 2006 JASIST on the top.

**Searching for intellectual turning points: Progressive knowledge domain visualization**  
By: Chen, CM (Chen, CM)

PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA  
Volume: 101 Pages: 5303-5310 Supplement: 1  
DOI: 10.1073/pnas.0307513100  
Published: APR 6 2004  
[View Journal Information](#)

**Abstract**  
This article introduces a previously undescribed method progressively visualizing the evolution of a knowledge domain's cocitation network. The method first derives a sequence of cocitation networks from a series of equal-length time interval slices. These time-registered networks are merged and visualized in a panoramic view in such a way that intellectually significant articles can be identified based on their visually salient features. The method is applied to a cocitation study of the superstring field in theoretical physics. The study focuses on the search of articles that triggered two superstring revolutions. Visually salient nodes in the panoramic view are identified, and the nature of their intellectual contributions is validated by leading scientists in the field. The analysis has demonstrated that a search for intellectual turning points can be narrowed down to visually salient nodes in the visualized network. The method provides a promising way to simplify otherwise cognitively demanding tasks to a search for landmarks, pivots, and hubs.

**Keywords**  
KeyWords Plus: AUTHOR COCITATION ANALYSIS; CO-CITATION; NETWORKS; DECOMPOSITION; GROWTH

**Author Information**  
Reprint Address: Chen, CM (reprint author)  
+ Drexel Univ, Coll Informat Sci & Technol, 3141 Chestnut St, Philadelphia, PA 19104 USA  
Addresses:  
+ [ 1 ] Drexel Univ, Coll Informat Sci & Technol, Philadelphia, PA 19104 USA  
E-mail Addresses: chaomei.chen@cis.drexel.edu  
+ Author Identifiers:

**Citation Network**  
104 Times Cited  
34 Cited References  
[View Related Records](#)  
[View Citation Map](#)  
[Create Citation Alert](#)  
(data from Web of Science™ Core Collection)

**All Times Cited Counts**  
113 in All Databases  
104 in Web of Science Core Collection  
23 in BIOSIS Citation Index  
8 in Chinese Science Citation Database  
0 in Data Citation Index  
1 in SciELO Citation Index

**Most Recent Citation**  
Mustafee, Navonil. Exploring the modelling and simulation knowledge base through journal co-citation analysis. SCIENTOMETRICS, MAR 2014.  
[View All](#)

Figure 63. The 2004 PNAS paper is cited 104 times, but the topic search won't be able to find it because the term CiteSpace does not appear in its title, abstract, or the keywords.

Now if you click on the Create Citation Report on the right, you will get access to all the records that citing this lot, i.e. that would be the citation expansion we want.

**Citing Articles: 66**  
(from Web of Science Core Collection)

For: Searching for intellectual turning points: Progressive knowledge domain visualization

Times Cited Counts  
113 in All Databases  
104 in Web of Science Core Collection  
23 in BIOSIS Citation Index  
8 in Chinese Science Citation Database  
0 data sets in Data Citation Index  
0 publication in Data Citation Index  
1 in SciELO Citation Index  
[View Additional Times Cited Counts](#)

Sort by: Times Cited -- highest to lowest

Page 1 of 7

Select Page | Save to EndNote online | Add to Marked List

1. **CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature**  
By: Chen, CM  
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY Volume: 57 Issue: 3  
Pages: 359-377 Published: FEB 1 2006  
[Get It](#) [View Abstract](#) Times Cited: 185 (from Web of Science Core Collection)

2. **Informetrics at the beginning of the 21st century - A review**  
By: Bar-Ilan, Judit  
JOURNAL OF INFORMETRICS Volume: 2 Issue: 1 Pages: 1-62 Published: 2000  
[Get It](#) [Full Text from Publisher](#) [View Abstract](#) Times Cited: 68 (from Web of Science Core Collection)

Analyze Results  
[Create Citation Report](#)

Figure 64. Citing articles to the 2004 PNAS paper.

The Citation Report shows, among other things, 732 citing articles. These 732 articles would form the expanded set. In fact, you can go even further by adding your search results to the [Marked List](#) ► [Create Citation Report](#) ► [Citing Articles](#). I will leave it to you to explore in the Web of Science.

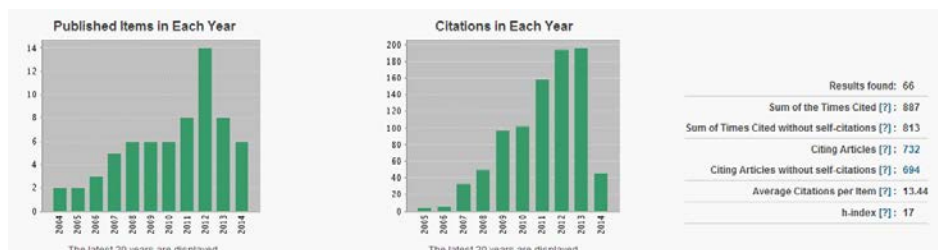


Figure 65. 732 Citing Articles will constitute the expanded set to download.

### 5.2.1.2 Download Records to Files

To download a set of records from the Web of Science, pull down the menu starting with Save to EndNote online and select Save to Other File Formats.

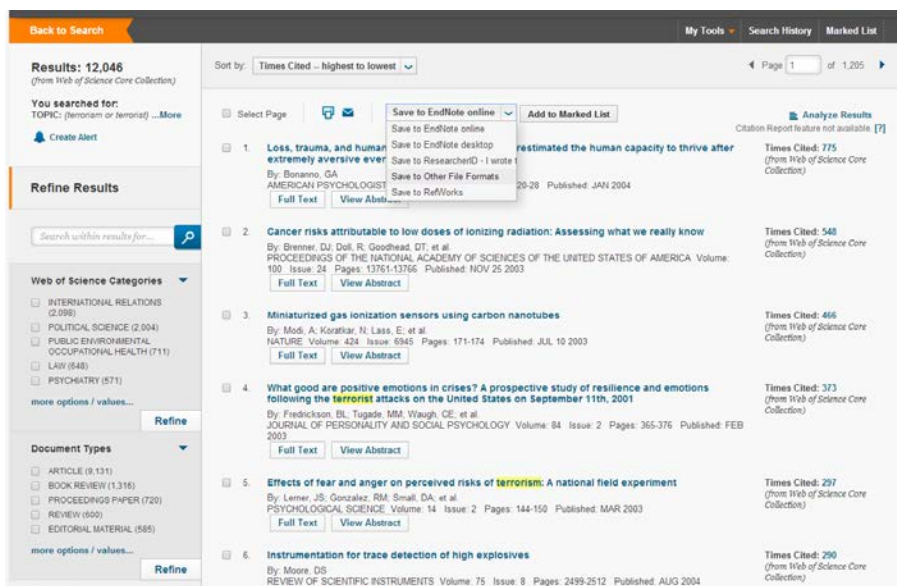


Figure 66. How to save records to other file formats.

Then you will need to enter the number of records, the content, and the file format in a dialog box like the following. For CiteSpace, include Full Record and Cited References and select Plain Text as the file format. When you save the file, make sure the file name starts with the word 'download' and the file extension is .txt. This naming convention will bring your more flexibility later on. For example, you can easily hide a file from CiteSpace by adding a prefix to the names of a few files you want CiteSpace to skip.

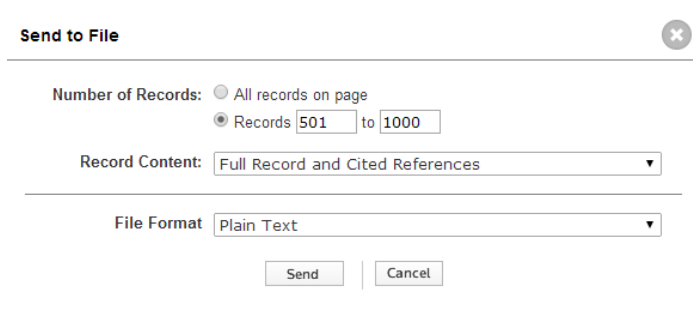


Figure 67. Download records 501-1000 in Plain Text. The Web of Science allows the maximum of 500 records each time to download. You may need to repeat the step multiple times.

## 5.2.2 Working with a CiteSpace Project

A CiteSpace project is designed to facilitate your analysis. Each project is associated with a dataset. You may analyze the dataset in many ways by selecting a variety of parameters and project properties. CiteSpace generates several types of intermediate files that you may want to inspect them in detail. You can handle most of these intermediate files directly.

### 5.2.2.1 Create a CiteSpace Project

You need to create two separate folders for a new project. One folder contains data files you just downloaded. We refer to it as the data folder. The other folder is the project folder, which will be used to store various intermediate files.

### 5.2.2.2 Edit an Existing Project

You can edit the properties of an existing project. To choose this function, full down the menu that shows “More Actions” next to the current project.

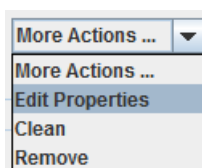


Figure 68. Edit the properties of an existing project.

You can edit several properties of an existing project based on your needs.

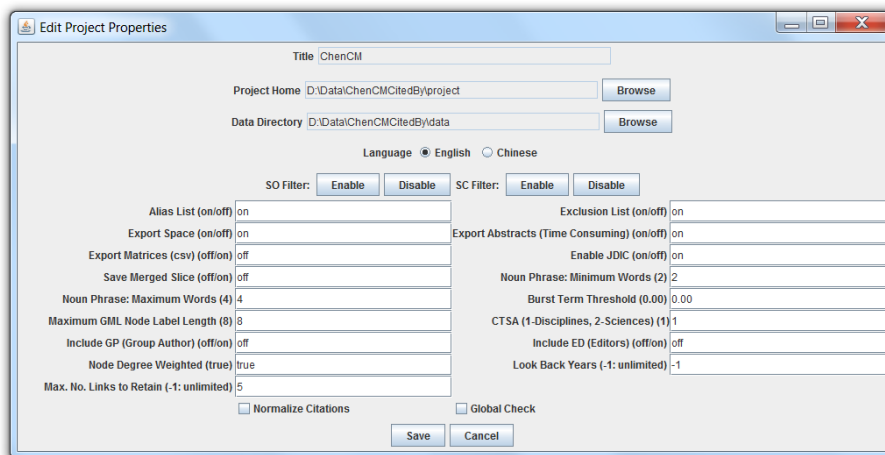


Figure 69. Properties of an existing project.

If you want to retain records from a specific set of journals in your dataset, you can enable the SO Filter function. First, you need to create a list of the names of journals in which those records you want to keep and save the list in an ASCII file as instructed below.

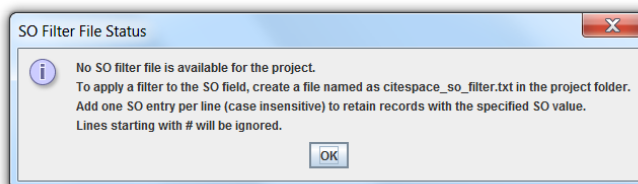


Figure 70. Instructions on creating an SO filter file.

You can similarly filter records based on their SC field, i.e. their subject categories.

#### Alias List: on/off

This property is used to enable or disable the feature of merging different variants of the same entity into a single node.

#### Exclusion List: on/off

This property is used to enable or disable CiteSpace to exclude a list of items to appear in the visualizations.

#### Look Back Years

This property controls the maximum length of a citation in terms of the difference between the publication dates of the citer and the cited reference. Set this property to -1 if you do not want any limit. For example, a value of 5 in this property means that citations made to references more than 5 years ago will be ignored.

This property is a simple link reduction method.

#### Max. No. Links to Retain

This property controls the maximum number of links to retain for each node in the network. Set this property to -1 if you do not want any limit.

For example, a value of 5 in this property means that up to 5 strongest links connecting to a node will be allowed. If the node has more than 5 connected neighbors, then they will be truncated, i.e. ignored.

This property is a simple link reduction method.

#### *5.2.2.3 Clean a Project*

This function will attempt to delete intermediate data files, for example, keyword extraction files, graph files in the graphml format, files of clusters, and files with the word *citespace* as the prefix of their filenames, which record how you configure your project.

CiteSpace will double check with you on some types of files to make sure you will not delete files that you may need.

#### *5.2.2.4 Remove a Project*

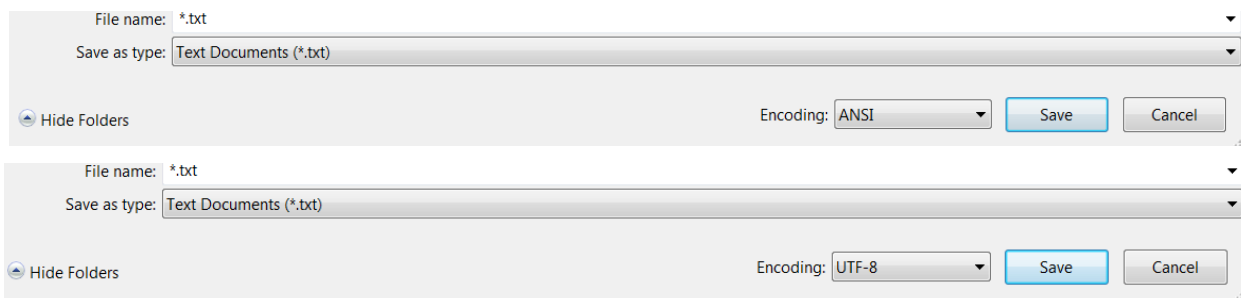
This function will remove the current project from CiteSpace, but it will leave the folders and files in these folders intact so that you can restore them by creating a new project and pointing to the existing folders.

### **5.2.3 Data Sources in Chinese**

A Java utility application that can convert data in the CSSCI format to the WoS format is available for download at the following link:

[http://cluster.ischool.drexel.edu/~cchen/citespace/utilities/CSSCIREC\(new\).jar](http://cluster.ischool.drexel.edu/~cchen/citespace/utilities/CSSCIREC(new).jar)

Store data files downloaded from CSSCI to a folder. Before using the format converter, make sure that the input data files with the ANSI encoding are saved to files with the UTF-8 encoding (Use any text editor and then Save As to files with the UTF-8 encoding). Then apply the converter to the data folder.



**Figure 71.** Save the downloaded CSSCI files in ANSI encoding to the UTF-8 encoding before using the format converter.

In order to use data files with Chinese encoding, use **Preferences ► Chinese Encoding**.

For more discussions in Chinese, see the following link:

<http://blog.sciencenet.cn/blog-496649-427780.html>

#### **5.2.4 How to handle search results containing irrelevant topics**

You may realize that no matter how carefully you formulate your search query in the Web of Science or any other sources, it is always possible that your search results contain irrelevant topics.

I recommend you to consider the following strategy. Instead of refining your query endlessly, you take the dataset that may include irrelevant topics and let CiteSpace to differentiate various topics. Until then, you should keep an open mind. You can determine whether a topic is indeed irrelevant only after you have a chance to examine the visualized results.

In most of the cases, it becomes straightforward to spot irrelevant topics because they would end up in an isolated cluster all by themselves. If it appears hard to tell, then the “irrelevant topics” may not be that irrelevant as you thought after all!

Here is an example to illustrate this point. The dataset was collected by a topic search for ‘hacker\*’, intended to catch topics relevant to hackers, hacker behavior, and associated topics.

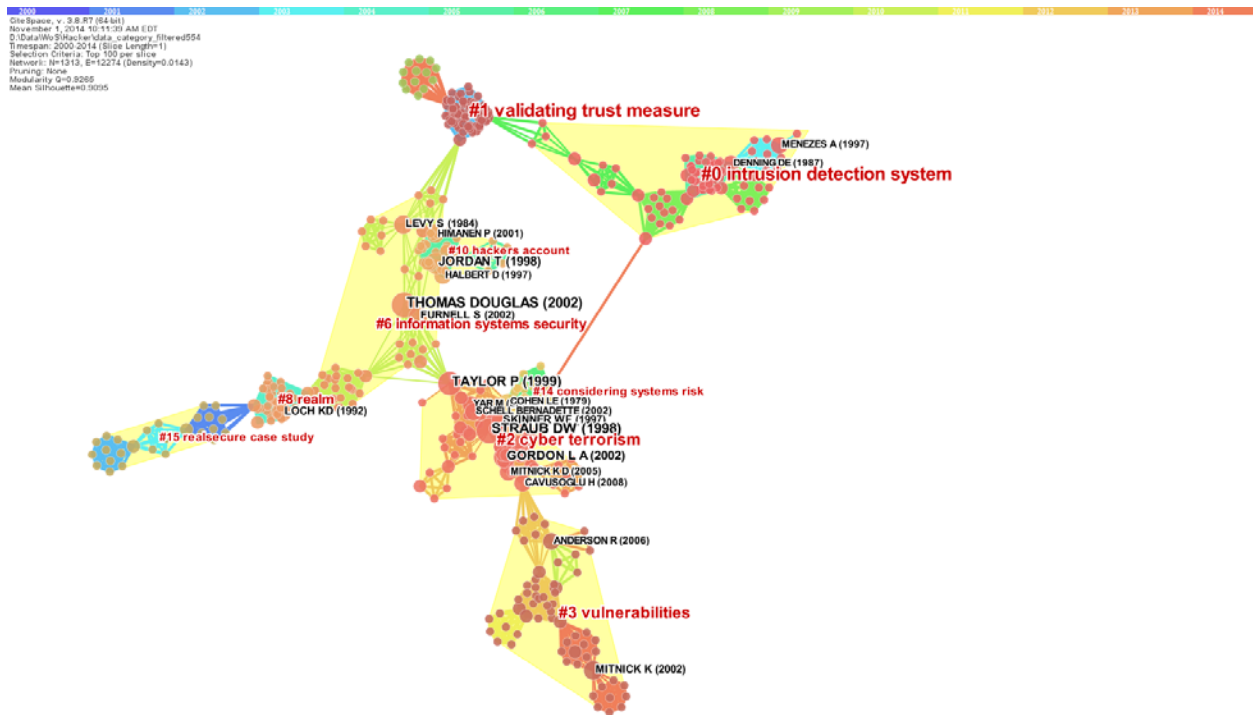


Figure 72. The largest connected component of a network of co-cited references.

The largest component shown above evidently contains relevant topics such as intrusion detection system, validating trust measure, cyber terrorism, and vulnerabilities. On the other hand, the visualization also reveals an interesting second largest component.

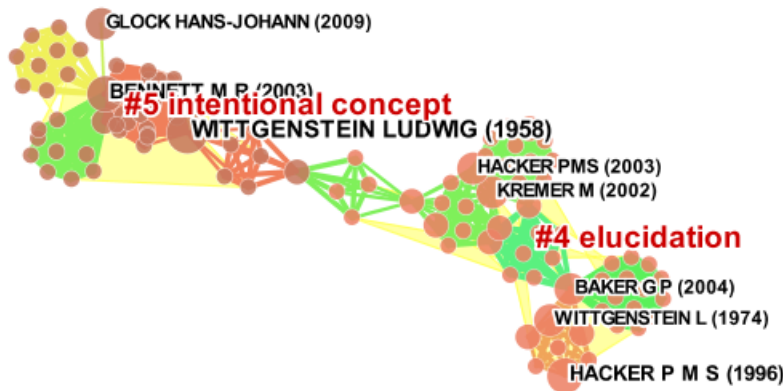


Figure 73. The second largest component of the network on 'hacker'.

The second largest component is not about the 'hacker' topics we wanted. Instead, these items were included in the original search result because the term Hacker is, as it turns out, the name of a prolific author – Hacker, P. M. S., a philosopher who published a number of articles in 1996 and 2003 and a book on Scepticism, Rules, and Language. The two clusters contained in the second largest component are essentially on topics irrelevant to the hackers in the context of computer security.

This example shows that it is a good idea to use a broader search query than a narrow one and defer the differentiation till the visual analytic process later on. The relevance of a topic would be much easier to detect at a later stage of the process.

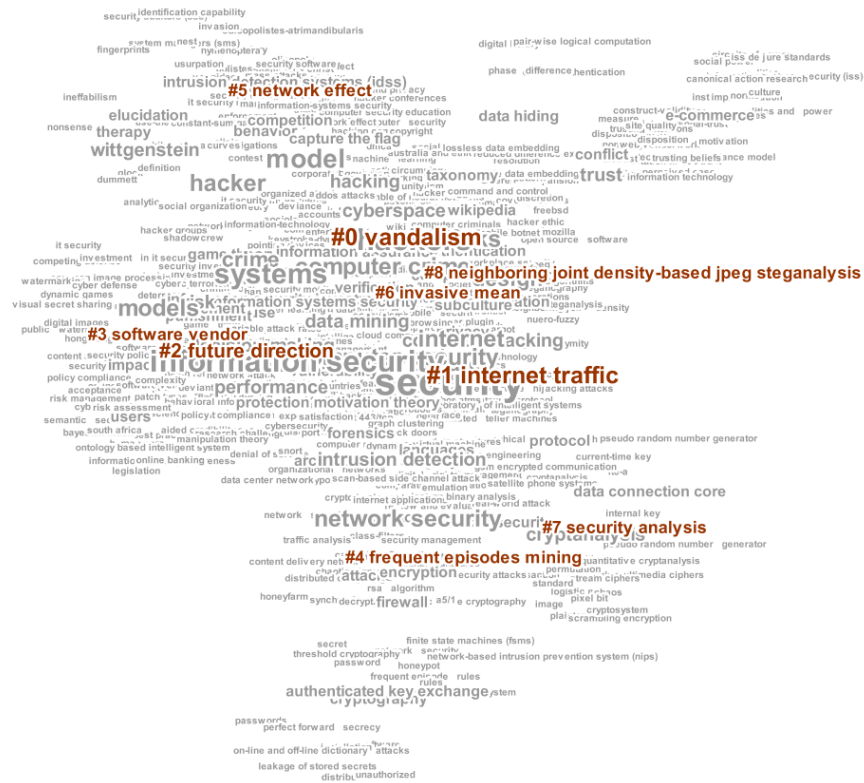


Figure 74. Clusters of a network of co-occurring keywords. Set Node size = 0 and Transparency = 0.

## 6 Configure a CiteSpace Run

A major process in CiteSpace is the network construction process. You can configure the process through a number of parameters. Your configuration will affect the results of the process.

### 6.1 Time Slicing

Given a dataset of bibliographic records, you need to choose the timespan that you want CiteSpace to analyze so that any records outside the timespan will be ignored. For example, your dataset may contain records from 1800s till 2014, you may choose to focus on the most recent 10 years or on a period in between. You can also include the entire dataset if you want to.

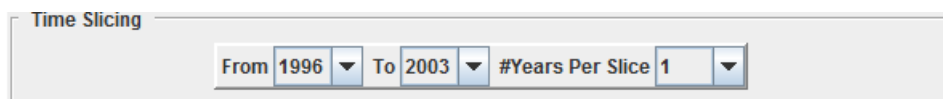


Figure 75. Configuring Time Slicing.

You can time slice the timespan in many ways by setting the value of #Years Per Slice. Typically, you would use 1-year slices and the number of networks will be the same as the number of years

within the timespan. Alternatively, you could use k-year slices so that each slide represents data of k years. You can also make a single slice so that you will only deal with one network.

The default selection is to divide the timespan into multiple 1-year slices.

## 6.2 Text Processing

Each bibliographic record contains four textual fields. These fields provide unstructured text that can be processed and analyzed as part of a visual analytic process.

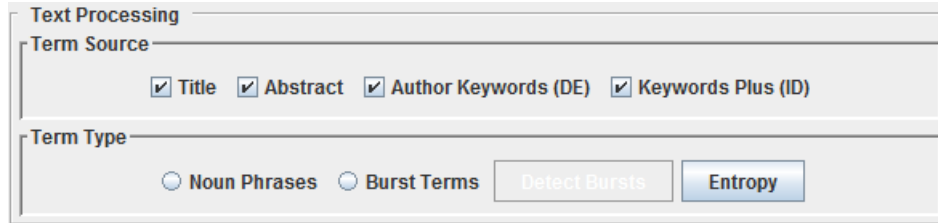


Figure 76. Settings for Text Processing.

You can skip the rest of this section if you are only interested in creating document co-citation networks, i.e. networks of cited references, or node types other than terms.

## 6.3 Configure the Networks

CiteSpace can generate several types of networks. The default node type is Cited References. In this case, the links are co-citation links. The networks are made of co-cited references.

CiteSpace allows you to choose a single node type or multiple concurrent node types. For example, you may select Author, Cited References, and Category to form networks of three types of nodes and 6 types of links, i.e. Author-Author (collaborative), Reference-Reference (co-citation), Category-Category (co-occurrence), Author-Reference (author-cites-reference), Author-Category (author-publishes-in-category), and Category-Reference (paper-in-category-cites-Reference).

Document co-citation networks are built on the methods pioneered by Henry Small (Small, 1973), but extended from a single-slide equivalent to multiple-slice network analysis, i.e. a time series of networks in order to detect critical transitions over time more effectively.

Author co-citation networks are originated from (White & Griffith, 1981).



Figure 77. Network Configuration.



Much of the attention in the design of CiteSpace has been devoted to document co-citation analysis due to the preferences that citation patterns of references provide particularly revealing insights into the structure and dynamics of scientific paradigms.

### 6.3.1 Bibliographic Coupling

If you choose Paper as the node type, the similarity between papers will be calculated by their bibliographic coupling (Kessler, 1963).

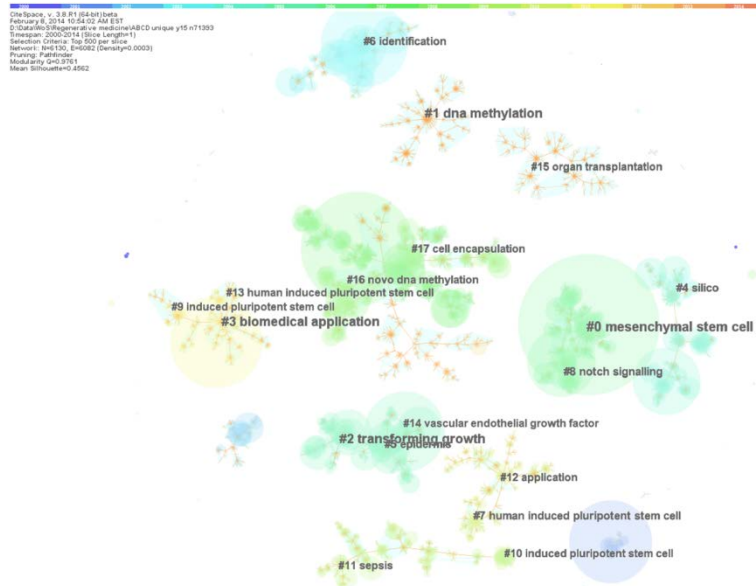


Figure 78. An example of a network of citing articles based on their bibliographic coupling scores.

Source: Chen, Dubin, and Kim (2014) *Emerging Trends and New Developments in Regenerative Medicine: A Scientometric Update (2000-2014). Expert Opinion On Biological Therapy. (In Press)*

## 6.4 Node Selection Criteria

CiteSpace provides several ways to sample records to form the final networks. These criteria are known as node selection criteria.

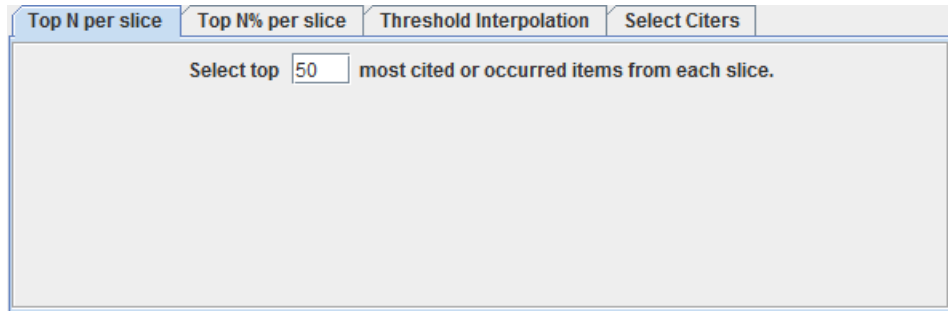
The simplest and recommended one is the first tab Top N per slice. If you enter a value of 50, then CiteSpace will select the 50 most cited or occurred items from each slice to construct a network, depending on the node types you selected in the previous step. If you selected multiple node types, then these nodes will be ranked by the number of times they appeared in the records for each slice.

The second selection method is Top N% per slice. For example, you can select the top 15% most cited items or most frequent items per slice. You can also select the entire dataset by specifying top 100% (as long as you raise the upper limit value high enough, say, 10,000 per slice).

The third method is Threshold Interpolation. It selects both nodes and links. It is complex. I recommend you to explore other selection criteria before this one.

The fourth one needs to be used along with one of the above 3 methods – Select Citers. You can select records based on a distribution of citations. You can specify an interval of the citation distribution, for example, an interval of [5, max] will include records that have 5 or more

citations. After the selection, you need to choose which one of the three selection methods you will need, namely, Top N, Top N%, or Threshold Interpolation.



**Figure 79. Node Selection Criteria.**

#### **6.4.1 Do I have the right network?**

Obviously the size of a visualized network influences the clarity and complexity of patterns we may learn from the visualization. The structure of a network is determined by the number of nodes selected for each time slice. It is unlikely that we will know in advance whether a Top N of 100 will generate a more desirable network than a Top N of 50.

Here are some suggestions:

First, begin with a Top N of 50 and generate a network visualization. Then check the modularity of the network, the number of clusters, and the average silhouette scores. We won't learn much from the network if there are only a couple of clusters. We won't get a big picture if there are hundreds of clusters either. A good range of the number of clusters would be about 7~10 major clusters with 10 or more members and each of the clusters has high silhouette values (e.g. > 0.70).

You can then try a Top N of 100 for each slice. If your computer is powerful enough, you can certainly try a Top N of 1,000 per slice or even higher.

You should start the process from a small network (although if you include many slices, even a Top N of 50 can accumulate to a large network), and then based on your initial assessment of the network enlarge the network accordingly.

Finally, note that the largest network is not necessarily the most informative one. Make clear the questions you want to answer first.

#### **6.5 Pruning, or Link Reduction**

Bibliographic networks can be very dense with many links. The process to remove excessive links systematically is called network pruning or link reduction.

CiteSpace provides two ways for this purpose: Pathfinder and Minimum Spanning Tree. A comparison of the pros and cons of the two methods is detailed in a 2003 publication (C. Chen & Morris, 2003). In a nutshell, Pathfinder is a theoretically better choice but it comes at a higher price.

I recommend you to start with networks without any pruning because sometimes pruning may reduce the characteristics of the natural groupings.

We are dealing with a time series of networks, i.e. sliced networks, and a merged network. When you select either Pathfinder or Minimum Spanning Tree, you will need to make another decision on whether you want to apply the pruning algorithm to all the individual sliced networks or the merged network only, or both. Since the merged network is resulted from what you do with the sliced networks, pruning sliced networks only will still lead to a merged network with reduced links. If you check both, then you will receive a merged network with the least number of links.



Figure 80. Pruning, or link reduction.

## 6.6 Visualization

By default CiteSpace will only show you the merged network. If you like, you can turn on the option to see networks of all the time slices. If you have 20 time slices, CiteSpace will open 20 extra windows for time sliced networks – you probably need to think twice before you do that!

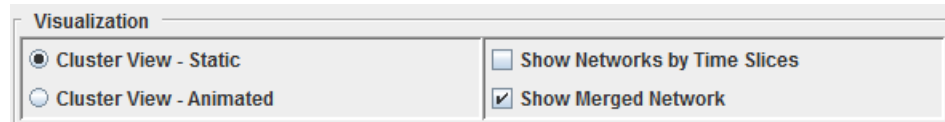


Figure 81. Visualization options.

## 7 Interacting with CiteSpace

Most controls of visual appearance are under the Display menu.

### 7.1 How to Show or Hide Link Strengths

To turn on/off the display of the strength of a link, use **Display ► Link Strength Show/Hide**.

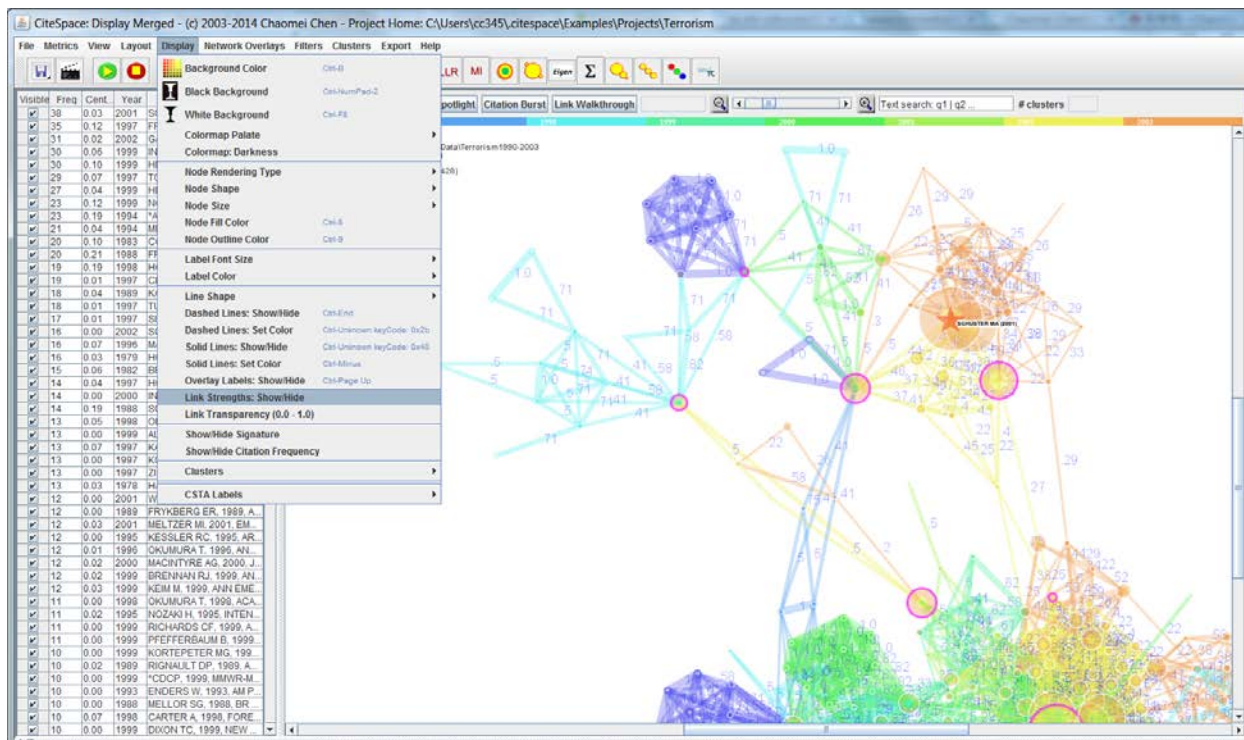


Figure 82. Showing the strength of each link in the display.

## 7.2 Adding a Persistent Label to a Node

In addition to labels controlled by the citation or frequency sliders, you can add a label to any node you like. Right-click on the target node and choose Label the Node.

To clear the label, right-click on the node again and choose Clear the Label.

Similarly, you can “bookmark” a node. A “bookmark” will show as a red star at the center of the node, like the one for the Schuster 2001 paper.

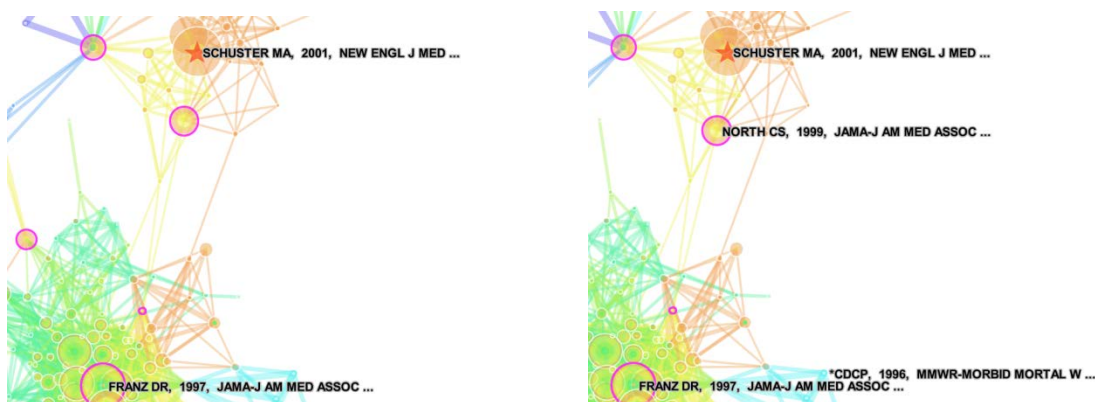


Figure 83. To add a persistent label to a node, right-click on the node and choose Label the Node.

### 7.3 Using Aliases to Merge Nodes

If you notice that some nodes in the network are in fact the variants of the same entity, you may use aliases to merge them so that they will appear as a single node. For example, in an author co-citation network below, CHEN CM and CHEN C are both from my own publications, so they should be merged into CHEN CM.

To use the alias function, first edit the properties of the current project and in particular make sure the Alias is on by typing an on and save.

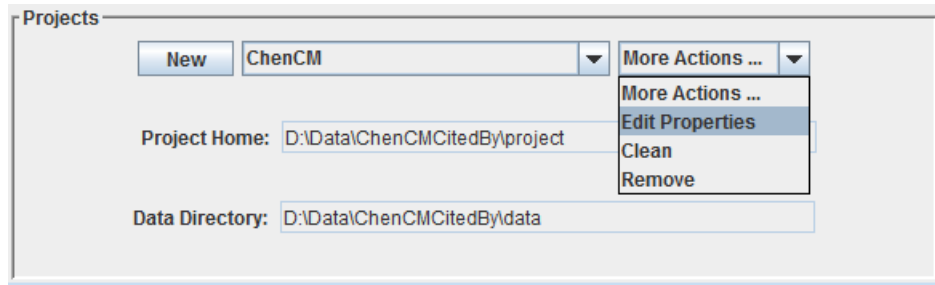


Figure 84. Edit the current project's properties.

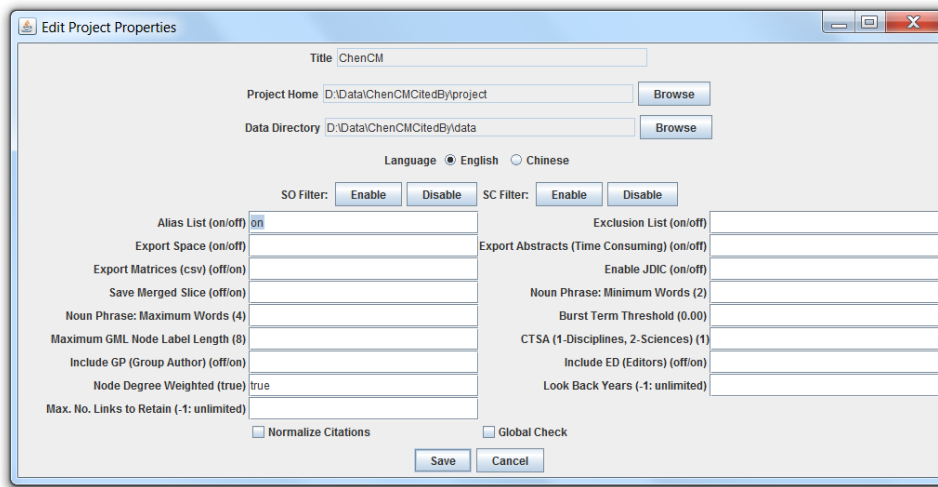
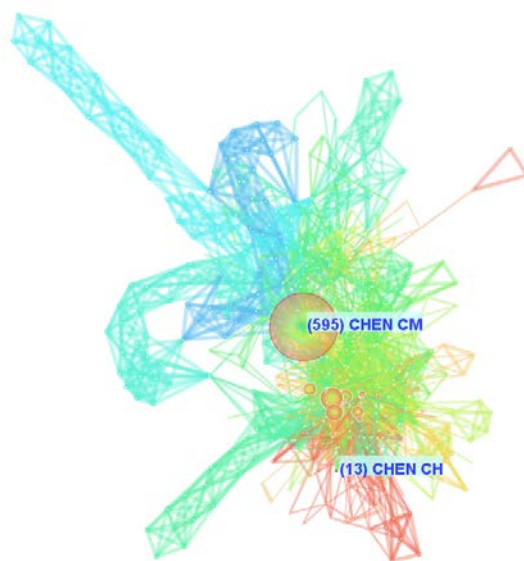


Figure 85. Make sure that the Alias List (on/off) is on. Type “on” in the field and save.

Right-click on the node CHEN CM and select it as the primary alias. Then right-click on the node CHEN C and select the secondary alias. CiteSpace will remind you that you need to re-run the process to see the changes.



**Figure 86.** Right-click on the node (574) CHEN CM and select “Add to the Alias List (Primary)” and select “Add to the Alias List (Secondary)” for the node “(133) CHEN C.”



**Figure 87.** The visualized network after the CHEN CM and CHEN C are merged.

In addition to merge nodes interactively through the graphical user interface, you can edit the `citespace.alias` file directly. The `citespace.alias` file needs to be located in the project folder for your target project. You can use any text editor to create and edit the file as long as the filename is `citespace.alias`.

The content of the file is formatted according to the following rules:

1. Each one contains a pair of node references separated by the ‘#’ character. The node references can be cited references, cited authors, or institutions.
2. The primary form of the alias, i.e. the node you want to retain, should appear first, and followed by the ‘#’ separator. The secondary alias, i.e. the node you want to merge into the primary alias node, should appear after the ‘#’ separator.
3. Include as many lines (i.e. pairs) as you need.

Save the file and go back to CiteSpace. Make sure the Enable Alias field in the Project setting is on. Next, you can start the project with GO!

The easiest way to learn about the various detailed formats you can use is to try out with a few examples through the interactive mode. Then open the CiteSpace generated `citespace.alias` file with a text editor to get yourself familiar with the existing example.

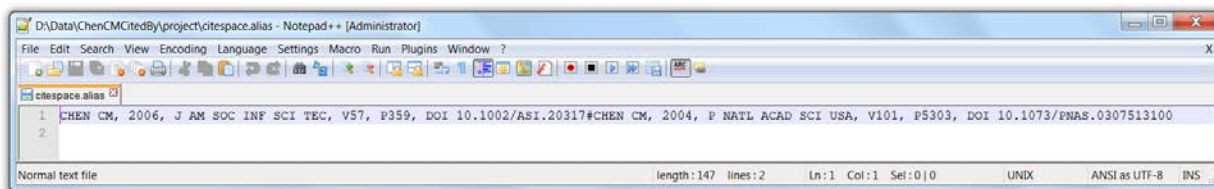


Figure 88. The content of the `citespace.alias` file is editable.

#### 7.4 How to Exclude a Node from the Network

You may exclude a node from the network by right-clicking on the node and select “Add to the Exclusion List”. You need to re-run the GO function to re-calculate the network model.

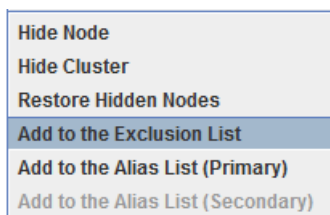


Figure 89. Exclude a node from the network modeling process as well as the visualization process.

The exclusion list is saved in the file `citespace.exclusion` in the project folder. You may edit it directly as a text file, for example, adding new entries directly or removing existing entries. If you want to remove the exclusion list altogether, simply rename or delete the `citespace.exclusion` file.

#### 7.5 How to Use the Fisheye View Slider

The fisheye view slider is provided for the timeline visualization so that you can see recent years are displayed with a larger screen estate than earlier years. As shown the in screenshots below, the majority of the publications are crowded in the recent few years in the original timeline visualization because some references are dated back as far as 1625, probably by philosophers.

Sliding the fisheye slider from 0 to 6 will help to spread out the crowded display so that we can each cluster’s activity in more detail than before.

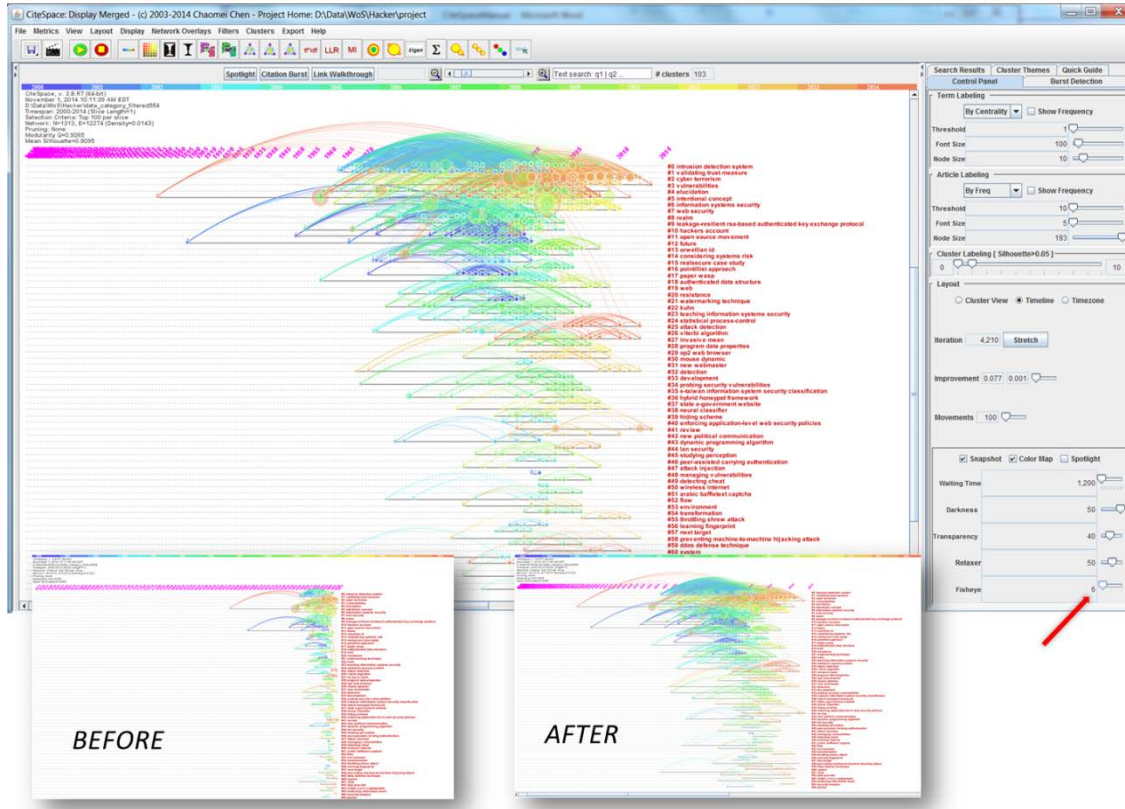


Figure 90. Use the Fisheye slider to adjust the timeline layout so that you can see recent years in more detail.

### 7.6 How to Configure When to Calculate Centrality Scores Automatically

As shown in the following figure, if the size of the network is more than 350 nodes, CiteSpace will turn off the automatic calculation of betweenness centrality scores in order to avoid an unnecessary delay.

```
Centrality(240): The calculation of centrality is deferred due to the size of the
network (653>350). Use CiteSpace->Preferences to reset the parameter.
Centrality(345): The network exceeds the centrality turn-off point (653>350). Use
CiteSpace->Preferences to reset the parameter.
```

Figure 91. CiteSpace turned off the automatic calculation of betweenness centrality scores.

If the auto-calculation is turned off, you will see all the values in the Centrality columns are 0s.

Visible	Freq	Cent...	Year	Cited References
<input checked="" type="checkbox"/>	38	0.00	2001	SCHUSTER MA, 2001, ...
<input checked="" type="checkbox"/>	35	0.00	1997	FRANZ DR, 1997, JAMA...
<input checked="" type="checkbox"/>	31	0.00	2002	GALEA S, 2002, NEW E...
<input checked="" type="checkbox"/>	30	0.00	1999	INGLESBY TV, 1999, JA...
<input checked="" type="checkbox"/>	30	0.00	1999	HENDERSON DA, 1999...
<input checked="" type="checkbox"/>	29	0.00	1997	TOROK TJ, 1997, JAMA...
<input checked="" type="checkbox"/>	27	0.00	1999	HENDERSON DA, 1999...
<input checked="" type="checkbox"/>	23	0.00	1994	*AM PSYCH ASS, 1994, ...
<input checked="" type="checkbox"/>	23	0.00	1999	NORTH CS, 1999, JAMA...
<input checked="" type="checkbox"/>	21	0.00	1994	MESELSON M, 1994, S...

Figure 92. The values in the third column Centrality are all 0s because centrality scores were not automatically calculated in this example due to the 653-node network is greater than the 350 cut-off point.



To start the centrality calculation manually, go to menu **Metrics ► Compute Centrality**.

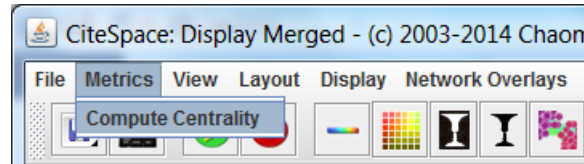


Figure 93. Manually start the centrality calculation.

Once the calculation is completed, you will see non-zero centrality values in the Centrality column.

Visible	Freq	Cent...	Year	Cited References
<input checked="" type="checkbox"/>	38	0.00	2001	SCHUSTER MA, 2001, ...
<input checked="" type="checkbox"/>	35	0.07	1997	FRANZ DR, 1997, JAMA-...
<input checked="" type="checkbox"/>	31	0.05	2002	GALEA S, 2002, NEW E...
<input checked="" type="checkbox"/>	30	0.05	1999	INGLESBY TV, 1999, JA...
<input checked="" type="checkbox"/>	30	0.07	1999	HENDERSON DA, 1999...
<input checked="" type="checkbox"/>	29	0.05	1997	TOROK TJ, 1997, JAMA-...
<input checked="" type="checkbox"/>	27	0.05	1999	HENDERSON DA, 1999...
<input checked="" type="checkbox"/>	23	0.11	1994	*AM PSYCH ASS, 1994, ...
<input checked="" type="checkbox"/>	23	0.06	1999	NORTH CS, 1999, JAMA...

Figure 94. The Centrality column now has non-zero values.

You can change the default threshold to disable the automatic centrality calculation. Go to menu **Preferences ► Set the Turn-Off Point of Centrality Computation** and enter a desirable number to the dialog box. For example, if you enter 1000, CiteSpace will use the new value next time to determine whether it will automatically calculate centrality scores or defer the calculation.

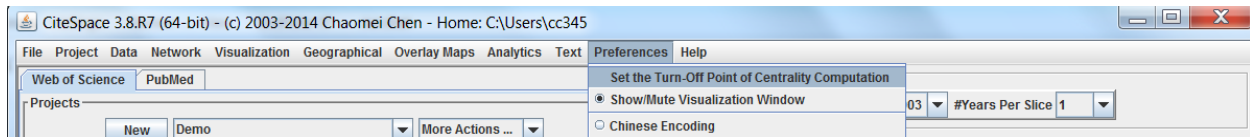


Figure 95. Set the threshold value.

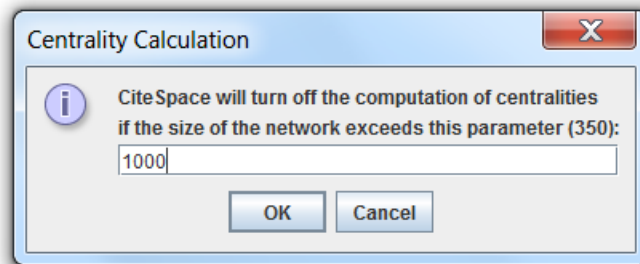


Figure 96. If you set the value to 1000, CiteSpace will automatically start to calculate centrality scores for networks with fewer than 1000 nodes.

### 7.7 How to Save the Visualization as a PNG File

You can save the visualization to a 300-dpi PNG file to the project folder on your computer. Click on the second icon under the menu bar.

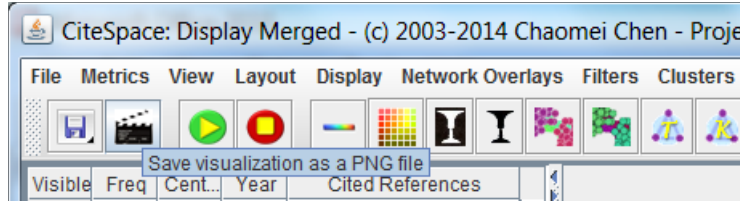


Figure 97. Click on the second icon to save the current visualization to a 300-dpi PNG file.

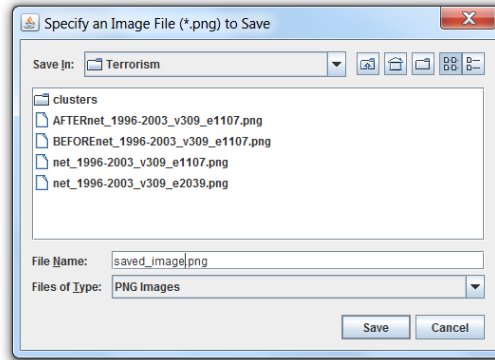


Figure 98. The default folder is the project folder of the current project. The default filename contains information about the network, i.e. the number of nodes is 309 and the number of edges is 2039.


Name	Date modified	Type	Size
 saved_image	11/10/2014 3:43 PM	PNG Image	930 KB

Figure 99. The new PNG image file is saved to your computer.

### 7.8 *Filters: Match Records with Pubmed*

If the topic you are analyzing is medical related, there is a good chance that you can find at least some of the records in PubMed. CiteSpace can build the bridge between a visualized network and corresponding records on PubMed by providing you with direct links to the display of these records on PubMed. On PubMed, you can explore further, for example, similar papers and other information.

The basic steps are as follows:

1. Using bibliographic records (with cited references) to generate a network visualization in CiteSpace
2. Divide the network into clusters and label these clusters as usual
3. Pull down the Filters menu and choose **Match records with PubMed**, and wait for the process to complete (since CiteSpace is set to comply with NCBI's protocol, it will tell you in advance how long you need to wait)
4. Right-click on a node or a cluster of your interest, choose **List Cluster Members** or **List Citers to the Cluster**

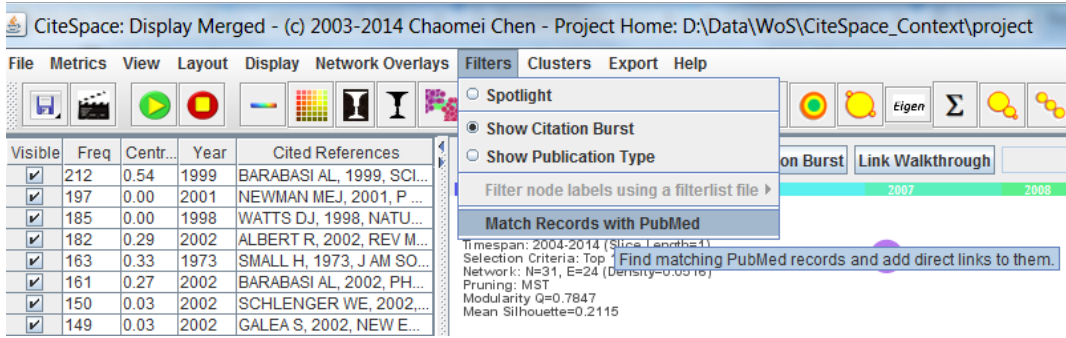


Figure 100. Choose the function **Match Records with PubMed** from the **Filters** menu.

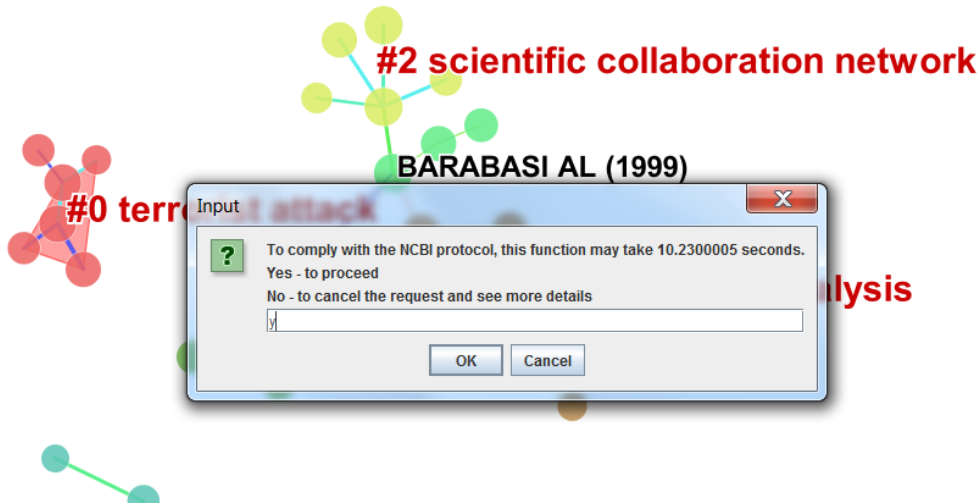


Figure 101. To comply with the NCBI protocol, CiteSpace estimates the time to be taken to complete the process.

```

PubMed(338): Searching for 31 records in PubMed ...
PubMed(383): 10% completed.
PubMed(383): 20% completed.
PubMed(383): 30% completed.
PubMed(383): 40% completed.
PubMed(383): 50% completed.
PubMed(383): 60% completed.
PubMed(383): 70% completed.
PubMed(383): 80% completed.
PubMed(383): 90% completed.
PubMed(389): Found 12 matches of the 31 records.
    
```

Figure 102. The progress of the process is reported in the command line prompt window. As you can see here, among the 31 records in the example, there are only 12 matches on PubMed. Therefore, bear in mind this will be only useful if the topic is not completely outside the scope of PubMed.

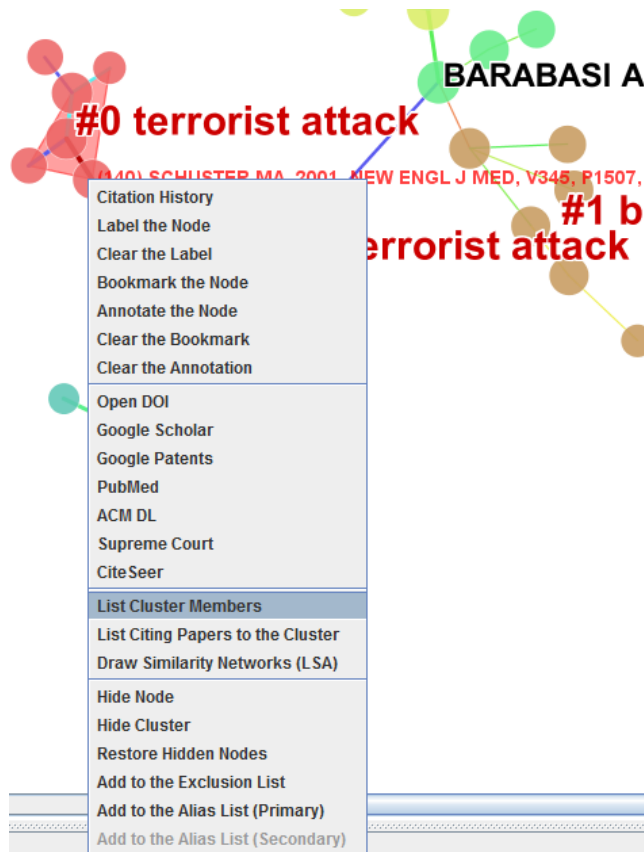


Figure 103. Right-click on a node in Cluster #0 terrorist attack (Schulster 2001) and select List Cluster Members.

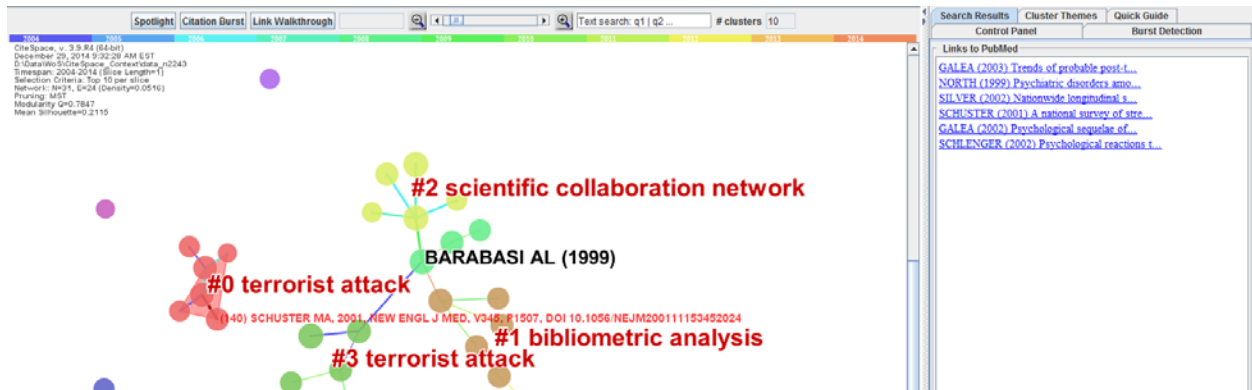


Figure 104. PubMed links of matched members of the cluster (#0) will be shown in the Links to PubMed panel under the Search Results tab.

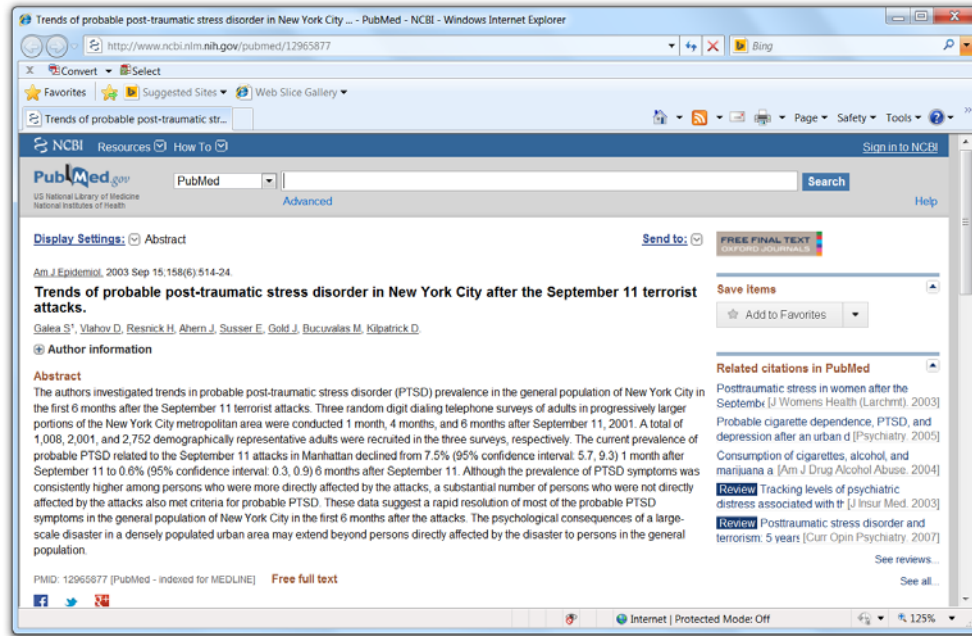


Figure 105. Click on a link in the **Links to PubMed** panel will bring you to the PubMed page in a new browser window.

- Citation History
- Label the Node
- Clear the Label
- Bookmark the Node
- Annotate the Node
- Clear the Bookmark
- Clear the Annotation

---

- Open DOI
- Google Scholar
- Google Patents
- PubMed
- ACM DL
- Supreme Court
- CiteSeer

---

- List Cluster Members
- List Citing Papers to the Cluster
- Draw Similarity Networks (LSA)

---

- Hide Node
- Hide Cluster
- Restore Hidden Nodes
- Add to the Exclusion List
- Add to the Alias List (Primary)
- Add to the Alias List (Secondary)

Figure 106. **List Citing Papers to the Cluster** will pop up a window that summarizes the citing papers of the cluster.

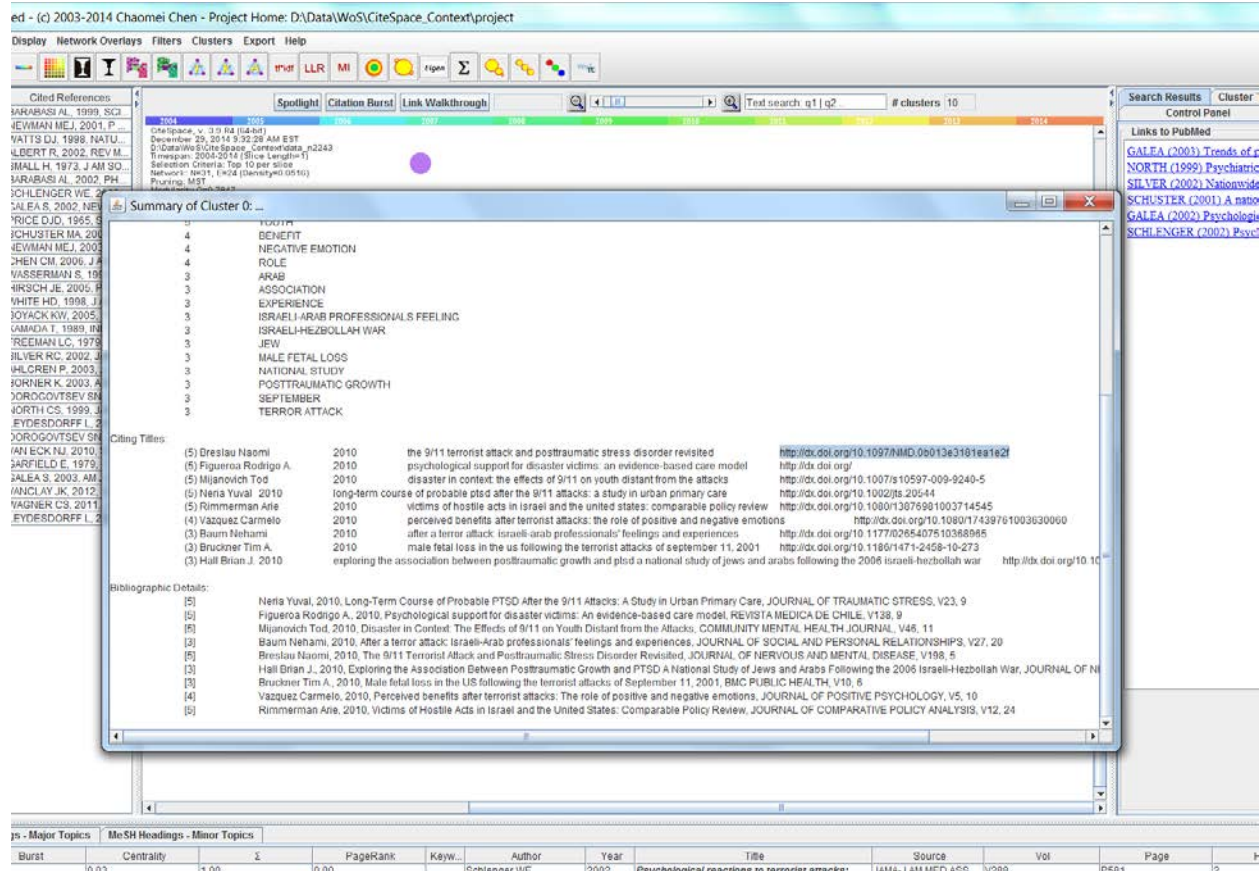


Figure 107. The Summary of the current cluster (based on where you right clicked), including keywords, citing papers, and their DOI links. You can copy the URL of a DOI link and paste it to a browser to get the full text if you have an adequate subscription.

## 8 Additional Functions

The main menu provides access to additional functions.

### 8.1 Menu: Data

Data ► Import/Export

CiteSpace provides some utility functions to facilitate data import and export needs.

#### 8.1.1 CiteSpace Built-in Database

CiteSpace provides a user interface to a MySQL database on localhost. The user interface provides various functions to import and export records in connection with the database.

Before you can use this group of functions, you need to set up your MySQL as follows.

On your computer, locate your own User folder and find the .citespace folder. Create a text file mysql.ini with the name-value pairs separated by a tab as the content:

```
host localhost
```

user *user\_id*  
pass *password*

where *user\_id* and *password* are your user id and password for your own MySQL login.

Make sure that your MySQL server is on before you use this function in CiteSpace.

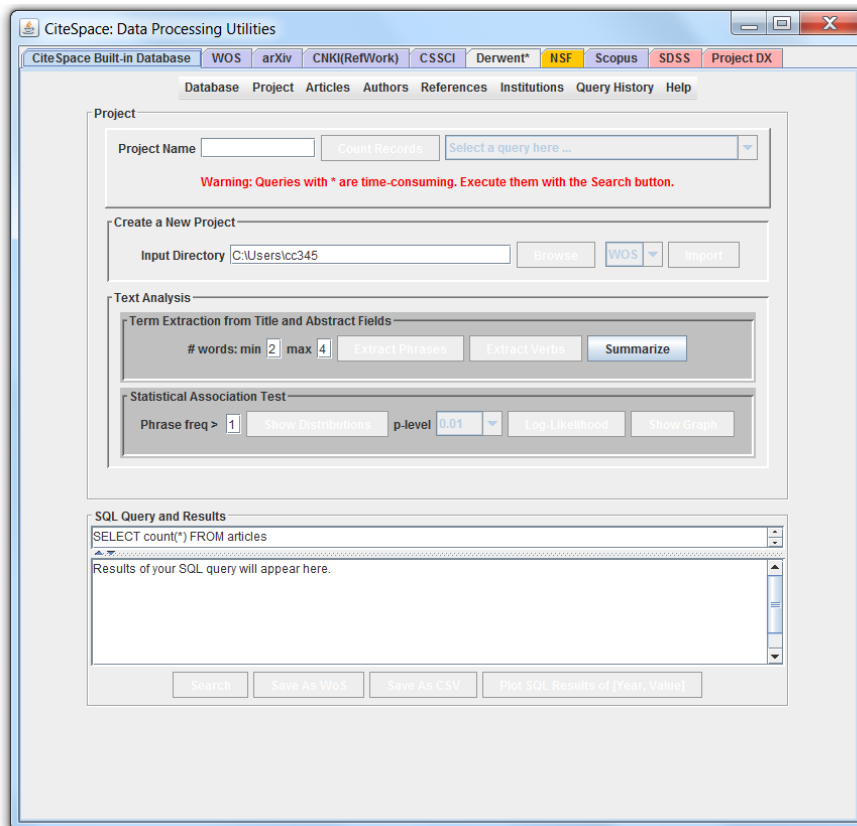


Figure 108. Data Processing Utilities.

After connecting to the database, you will see existing projects, i.e. projects that have been loaded to the database. Note the projects here are stored in the database and they are different from the projects appeared on the main interface of CiteSpace, which are file-based, i.e. the files you downloaded from the Web of Science. You can import the downloaded files to the database and edit them accordingly and export to files in the Web of Science format.

Since the database is a MySQL database on localhost, you can access the database directly with your own MySQL login. You can use this database to process your data before you apply visualization functions on them.

#### 8.1.1.1 Structure of the Database

The name of the database is wos. It contains the following tables:

TABLE articles

id(int), uid, project, author, title, abstract, source, j9, volume, issue, bp, ep, page, dt, doi, year(int), month(int), date(int), citations(int), editor, tagged(boolean)

TABLE authors

id(int), lastname, firstname, initials, project, uid, pos

TABLE refs

id(int), bibcode, ref, doi, author, year, source, volume, page, citer\_uid, project

TABLE keywords

id(int), keyword, uid, year, project, type

TABLE phrases

id(int), phrase, isTitlePhrase(booean), project, uid, year(int), month(int), date(int), freq(int)

TABLE verbs

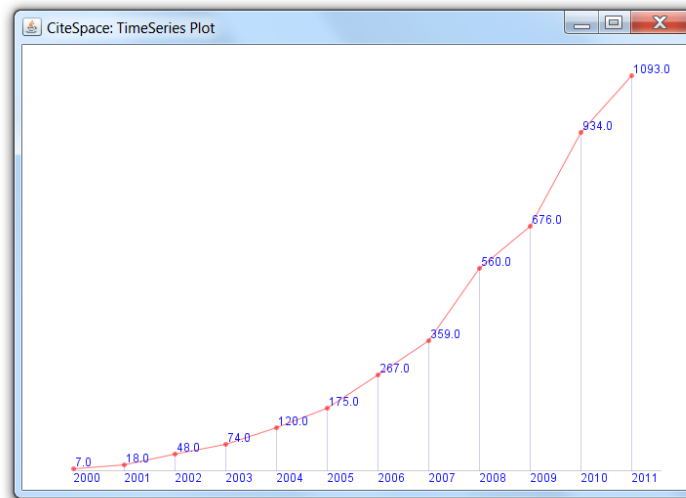
id(int), project, uid, verb, freq

TABLE bursts

id(int), project, term, weight(double), start(int), end(int)

TABLE institutions

id(int), name, country, uid, year(int), project



**Figure 109.** A plot from a project in the built-in database.

### Articles ► Most Cited Articles

You can query the database with a few built-in functions on a loaded dataset. For example, you can find the most cited articles in the current project. The SQL query is displayed along with the results. It will help you to get familiar with the internal structure of the database.



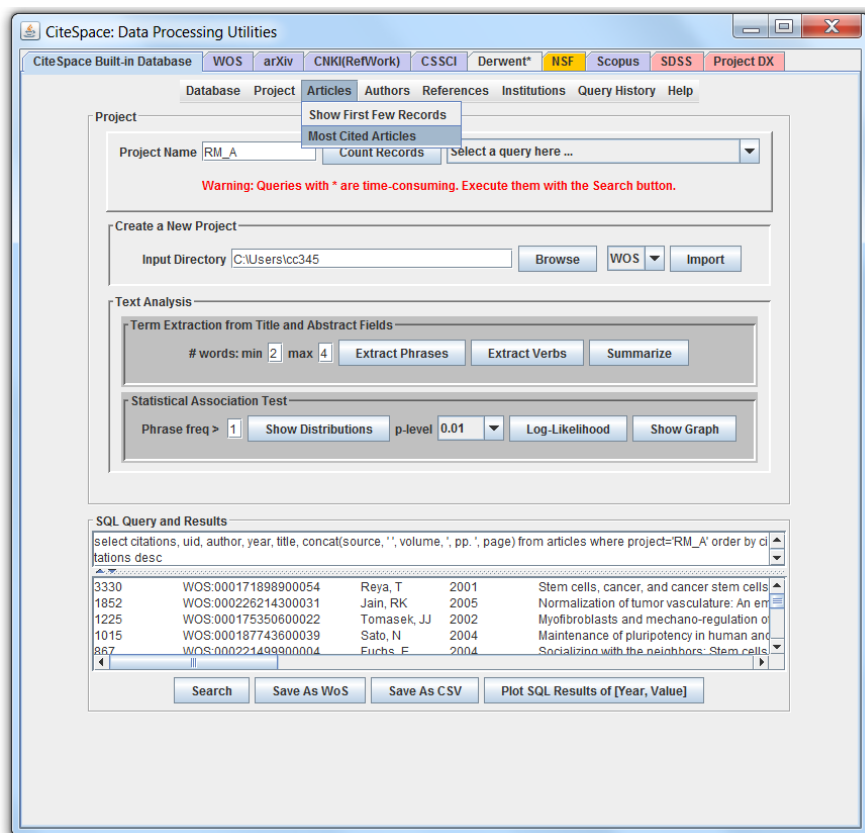


Figure 110. Using a built-in function to find the most cited articles with a SQL query.

## 8.1.2 Utility Functions for the Web of Science Format

### 8.1.2.1 Removing Duplicate Records

You can merge multiple datasets you have downloaded by merging the downloaded files to the same data folder. If some files have the same names, you will need to rename them first to resolve the conflicts before you move them together. The simplest way is to add a suffix to the names of the files. For example, if you have two datasets and each contains a file named download\_500.txt, you can rename them to download\_500\_part1.txt and download\_500\_part2.

You will need to make sure that the merged files do not have duplicated records. CiteSpace has a utility function for this. Specify the input folder and the folder to save a copy of the dataset after duplicates are removed, then press the button “Remove duplicates (WoS)”. Note the format of the input files must be in the Plain Text format of the Web of Science.

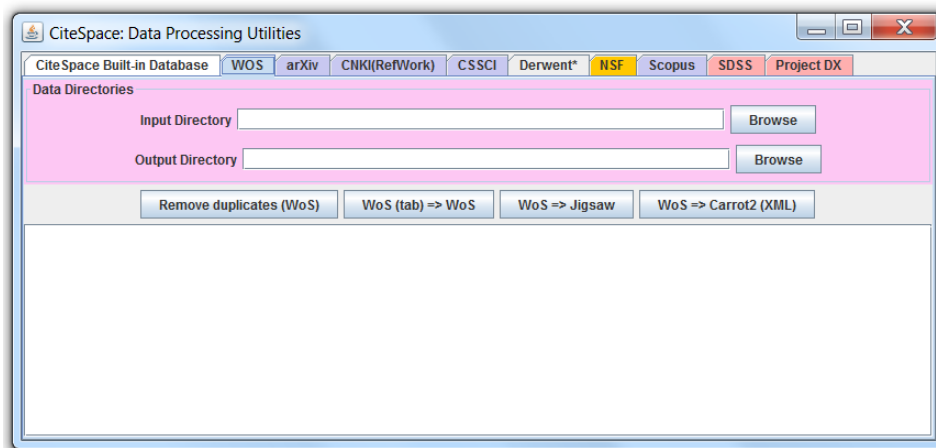


Figure 111. Utility functions for handling bibliographic records in the Web of Science format.

#### 8.1.2.2 Convert the Tab Delimited WoS Format

You can convert the tab delimited WoS format to the Plain Text format (i.e. each field is marked by a two-letter code such as AU, TI, and AB) using another utility function “WoS(tab) → WoS.”

#### 8.1.2.3 Convert the WoS Format for Jigsaw

You can convert files in the WoS format to a format that can be processed by Jigsaw – a visual analytic application, which is also freely available (Stasko, Gorg, & Liu, 2008).

#### 8.1.2.4 Convert the WoS Format for Carrot2

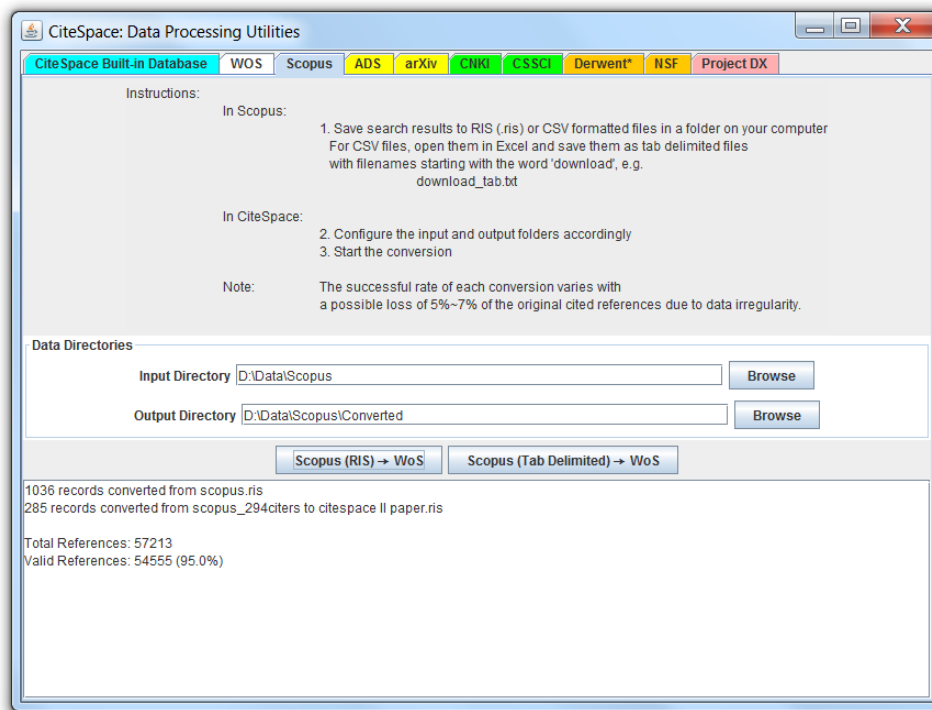
You can convert files in the WoS format to a format that can be processed by Carrot2 – an open source text search and visualization tool (“Carrot2: Open source framework for building search clustering engines,” 2012). The converted files are XML documents.

### 8.1.3 Scopus

The preferred format for Scopus data files is the RIS format. In Scopus, save your search results to one or more data files in RIS, e.g. `my_scopus_search_results.ris`, to a folder on your computer. Specify the data folder and a folder where you want to have the converted files. And start the conversion process.

CiteSpace will tell you how many records in each RIS files have been converted and, more specifically, how many cited references in total are found in the data files and how many of them have been converted successfully. The 95.0% is a very decent successful rate, considering all the irregularities of the cited references.

If, for some reason, you have the Scopus data files in CSV instead of in RIS, you need to do a quick conversion from CSV to a tab delimited format using Excel before you can run the Tab Delimited Converter in CiteSpace.



**Figure 112. Converting Scopus data files to the WoS format for CiteSpace.**

Name	Date modified	Type	Size
download_converted_1	12/29/2014 11:22 ...	TXT File	4,658 KB
download_converted_2	12/29/2014 11:22 ...	TXT File	1,176 KB

**Figure 113. Converted Scopus data are saved in the designated output folder.**

### 8.1.4 PubMed

CiteSpace allows you to retrieve bibliographic records from PubMed. For example, to retrieve records on hypertension based on MeSH headings you can use the query “hypertension [mh]” between 2008 and 2011. You can specify the maximum number of records you want to retrieve each year. For illustrative purposes, we limit the maximum number to 25 per year. Retrieved records will be saved to a special folder \$your\_username\PubMed\SearchResults.

Once the data retrieval is completed, you need to switch to the Web of Science tab and analyze the data in the same way as you did with a dataset from the Web of Science.

Since PubMed records do not include information on cited references, it is not possible to perform citation analysis, i.e. you cannot choose the node types such as cited references, cited authors, or cited journals. Nevertheless, you can perform other analyses such as networks of collaborative authors, terms, keywords, and categories.

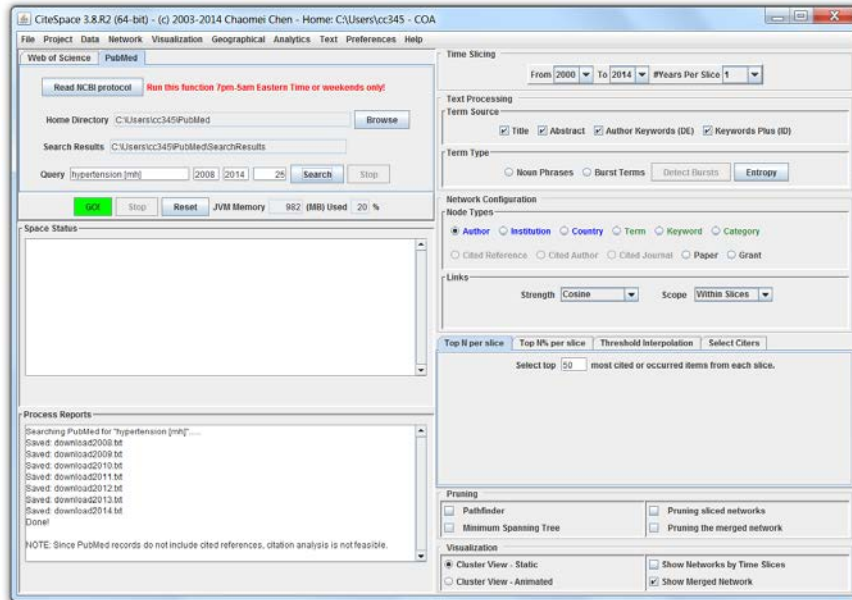


Figure 114. Retrieve bibliographic records from PubMed.

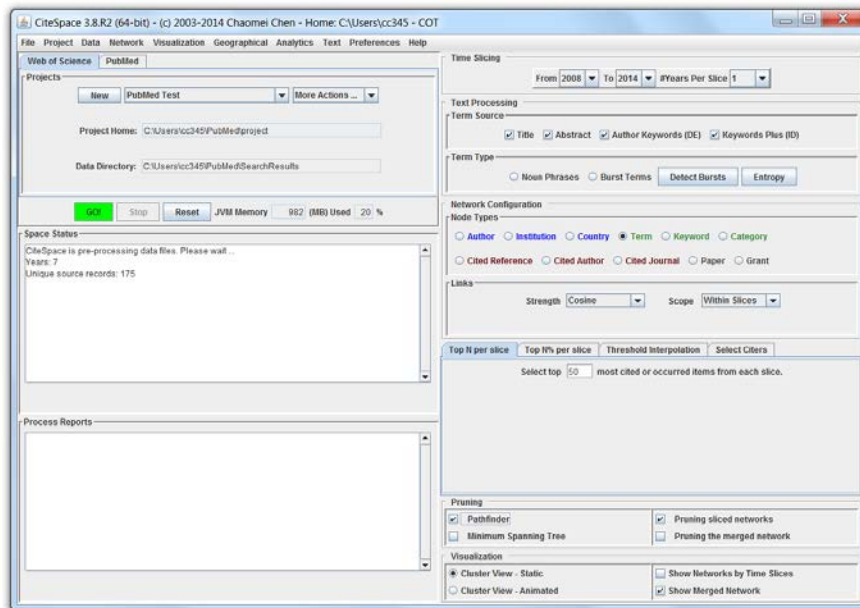


Figure 115. Analyzing the PubMed records ...

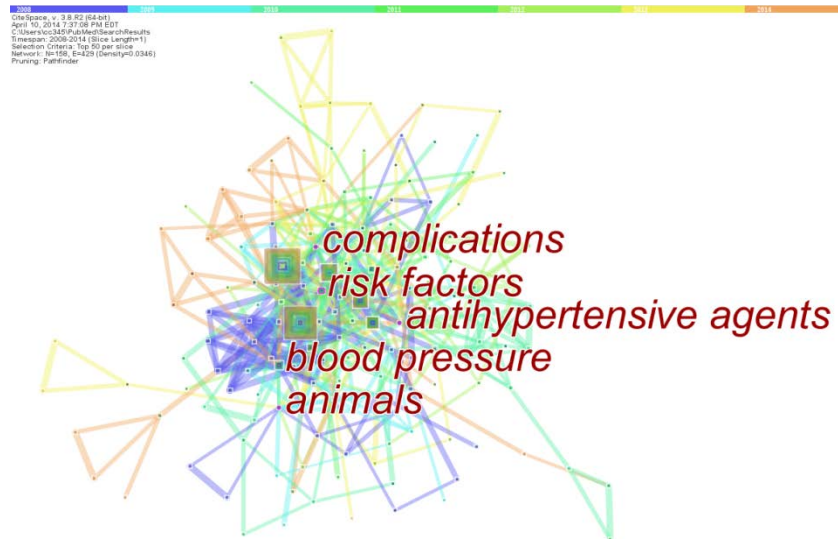


Figure 116. A network of co-occurring noun phrases on hypertension.

## 8.2 *Menu: Network*

### 8.2.1 Batch Export to Pajek .net Files

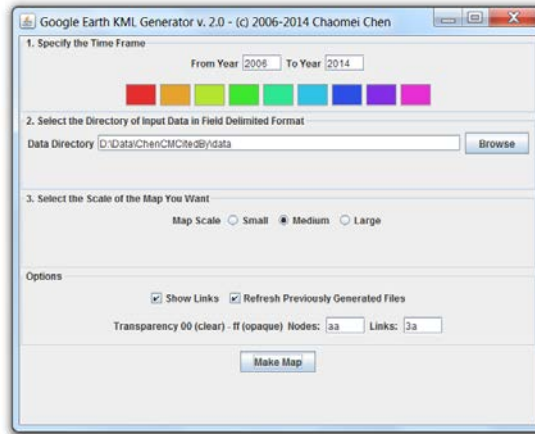
## 8.3 *Menu: Geographical*

### 8.3.1 Generate Google Earth Maps

Authors' geographic locations in their publication records can be mapped to a geospatial map in KML. You can use Google Earth as the interface to explore the authors' locations and links to their collaborators. You can also go to the original articles directly within Google Earth.

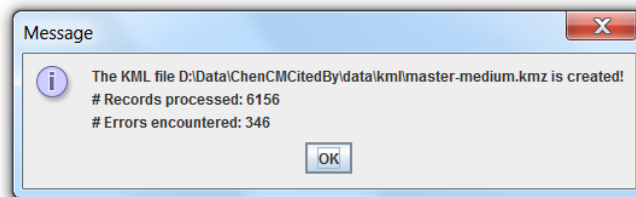
To generate the map file, you need to specify a data folder that contains bibliographic records in the Web of Science format (plain text), which is the same format for CiteSpace projects. This time we just need the data folder. A new folder will be automatically created under the data folder called kml. You will find the generated KML file in the kml folder when the geocoding process is completed.

The Google Earth map generator from CiteSpace needs to know the timespan you are interested, similar to the time slicing setup in the main interface of CiteSpace. Browse to the data folder of your data and click on the "Make Map" button. It may take a while for the process to complete.



**Figure 117. Google Earth KML Generator.**

Once the map is generated, you will see a Message notifying you where the map file is, which is in kmz format, i.e. a compressed KML file.



**Figure 118. The map is generated.**

If you see some errors reported by the generator, you may check the error log file – `geocoding_log_tab.txt` – and see if you can make corrections in the data and repeat the process afterwards. The map is stored in the `master-medium.kmz` file if you use the default scale of medium.

Name	Date modified	Type	Size
geocoding_log_tab	4/12/2014 10:55 PM	TXT File	47 KB
locations-2006	4/12/2014 10:53 PM	Microsoft Excel Co...	6 KB
locations-2007	4/12/2014 10:54 PM	Microsoft Excel Co...	8 KB
locations-2008	4/12/2014 10:54 PM	Microsoft Excel Co...	11 KB
locations-2009	4/12/2014 10:54 PM	Microsoft Excel Co...	16 KB
locations-2010	4/12/2014 10:55 PM	Microsoft Excel Co...	14 KB
locations-2011	4/12/2014 10:55 PM	Microsoft Excel Co...	21 KB
locations-2012	4/12/2014 10:55 PM	Microsoft Excel Co...	17 KB
locations-2013	4/12/2014 10:55 PM	Microsoft Excel Co...	15 KB
locations-2014	4/12/2014 10:55 PM	Microsoft Excel Co...	5 KB
master-medium	4/12/2014 10:55 PM	KMZ File	59 KB

**Figure 119. The generated files in the KML folder.**

If you have Google Earth installed on your computer, you can double click on the kmz file.

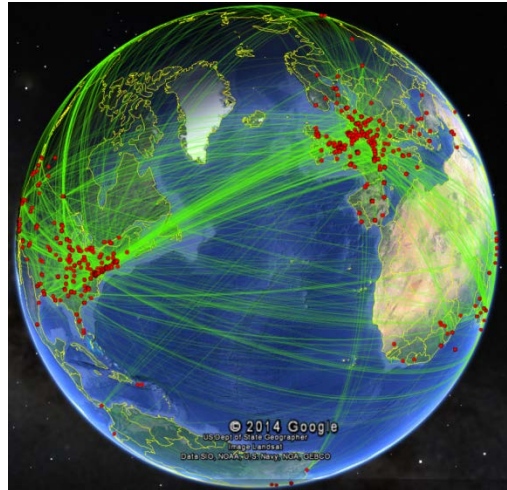


Figure 120. The author collaboration network is shown in Google Earth.

Under the Places, you will see a list of years as layers. You can select or unselect these layers by checking or unchecking the checkbox in front of them so that you can control which years of data you want to see. Coauthored papers in more recent years are linked by lines in red, whereas older collaborations are shown in green or blue lines.

You can drill down from a layer of a year to a location, then to a list of papers published by authors at that location. Each paper on the list is clickable. It will bring you to its full text via its DOI link. You need to have the right subscription to access papers in this way.

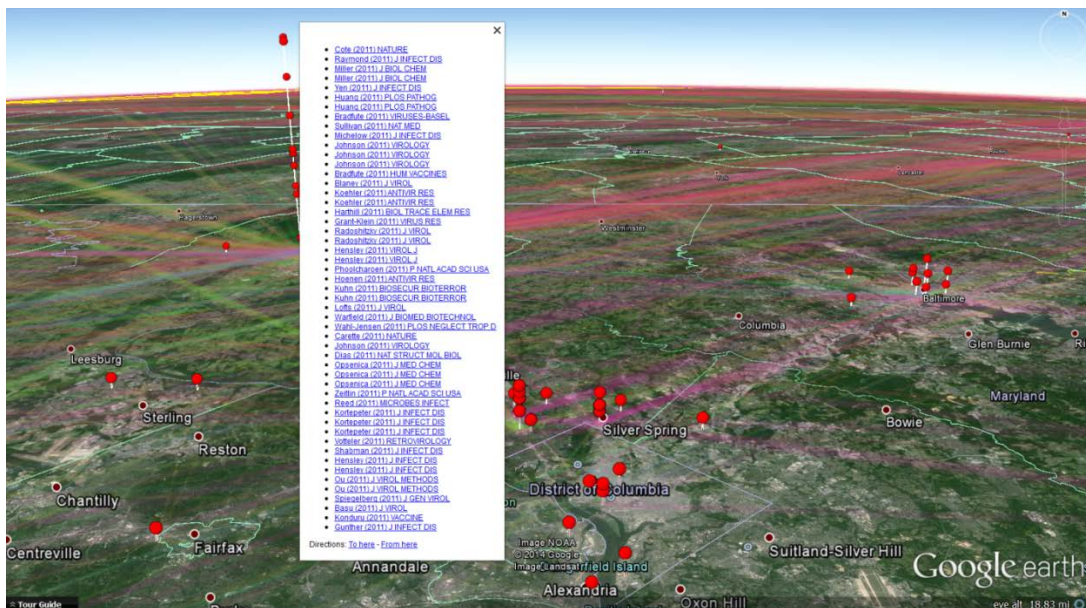


Figure 121. Unfold the list of places on the left and locate a city of interest – Frederick, Maryland, USA – on Ebola.

Click on any of the papers on the list to explore its content. Here is an example of what you will see after clicking on a link to our 2010 JASIST paper in Google Earth.

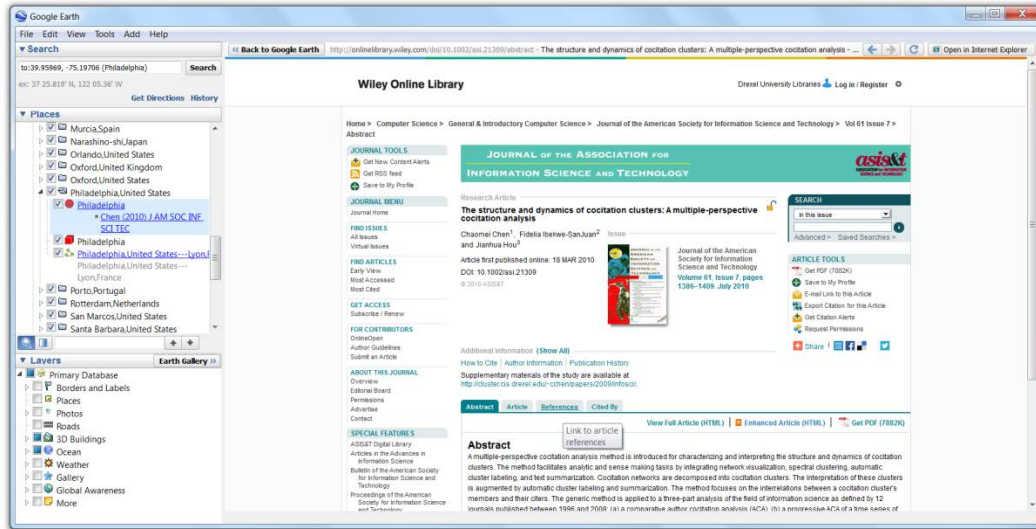


Figure 122. Clicking on the Chen (2010) link takes us to the publisher's page of the paper.

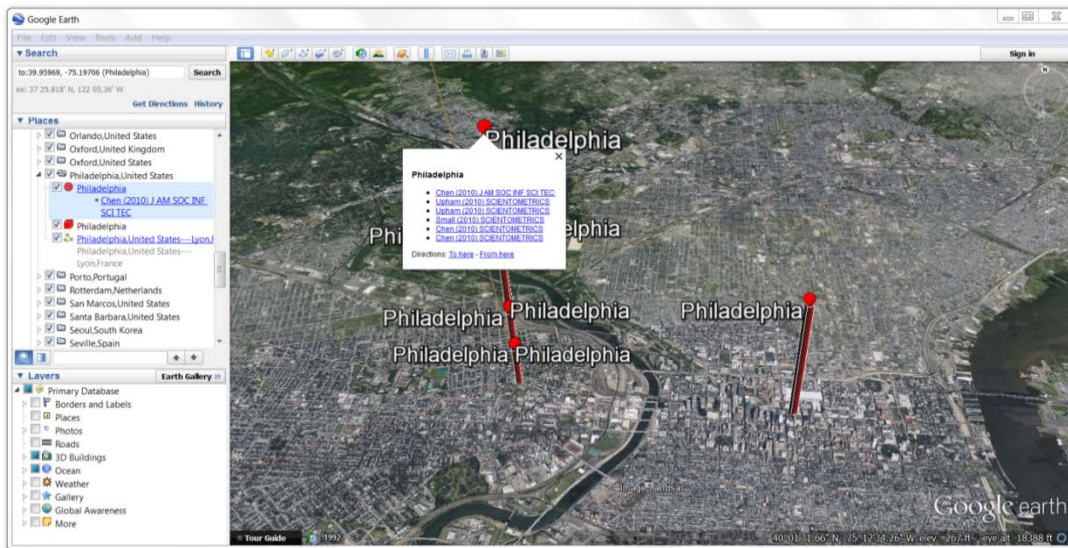


Figure 123. Here is a bird eye view of the downtown Philadelphia. The red bar on the left is on Drexel's main campus.

#### 8.4 Menu: Overlay Maps

Dual-map overlays are introduced in (C. Chen & Leydesdorff, 2014).

This function is made available for non-commercial and educational use. For commercial use, please contact me at [chaomei.chen@drexel.edu](mailto:chaomei.chen@drexel.edu) directly for further detail.



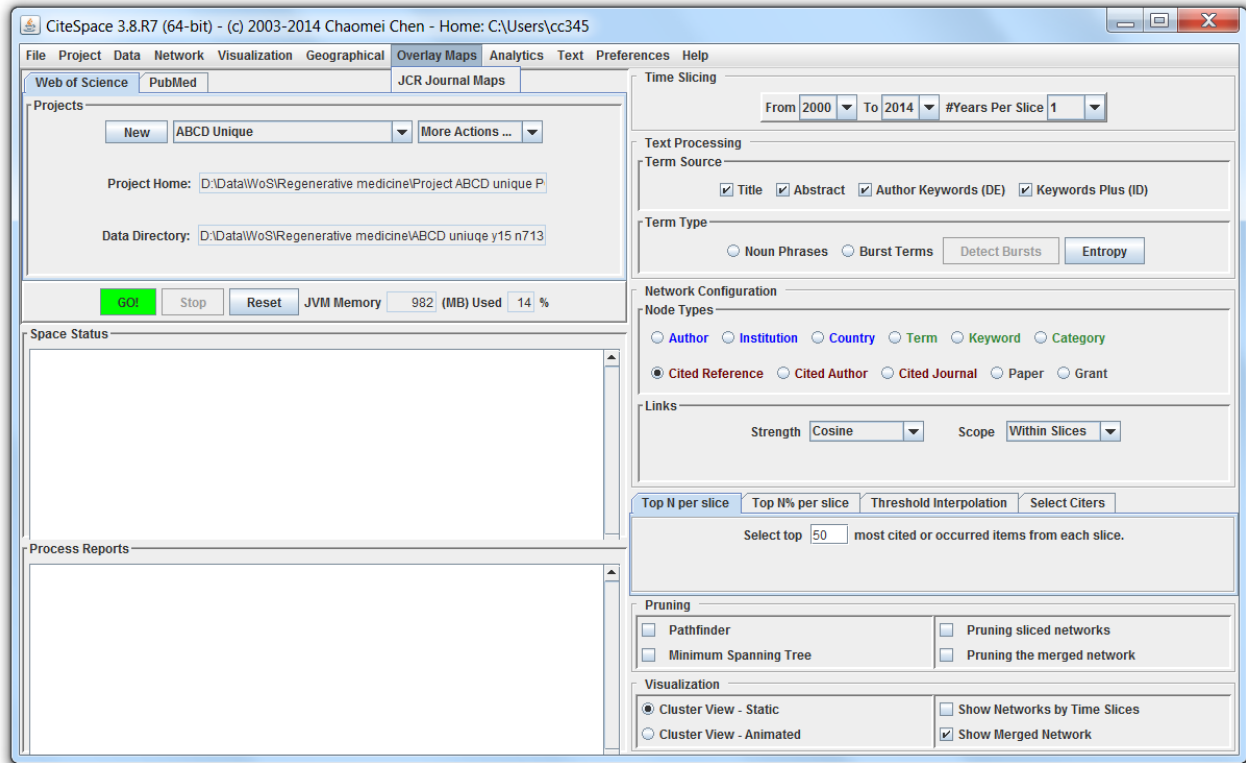


Figure 124. The dual-map overlay maps function is accessible from the Overlay Maps menu.

### 8.4.1 Add an Overlay

The current version allows you to add up to 12 overlays. Each overlay is represented by a distinct set of bibliographic records in the WoS format. Start the process by clicking on the Add button.

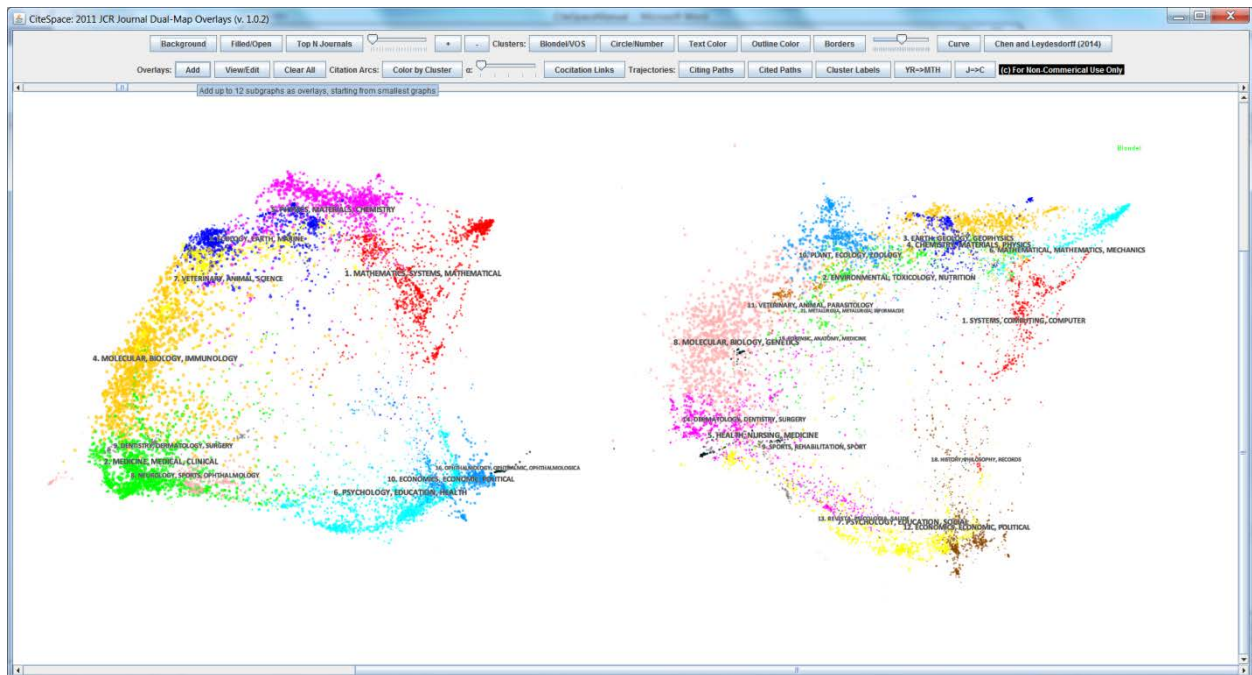


Figure 125. Start the process of adding an overlay by clicking on the Add button.

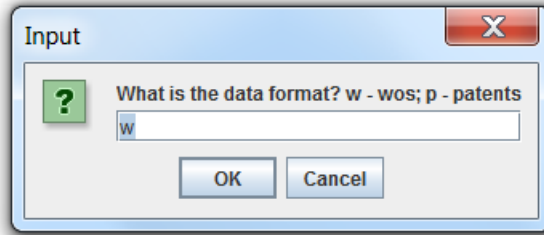


Figure 126. Select 'w' for bibliographic records in the WoS format. The 'p' option is not available in the current release.

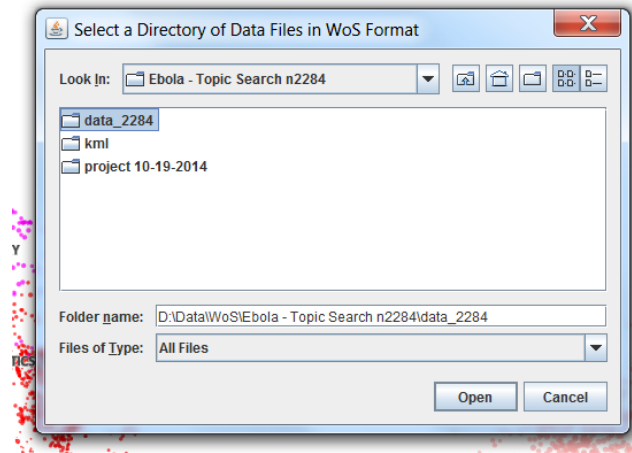


Figure 127. Select the data directory of a dataset in the same way as you create a CiteSpace project.

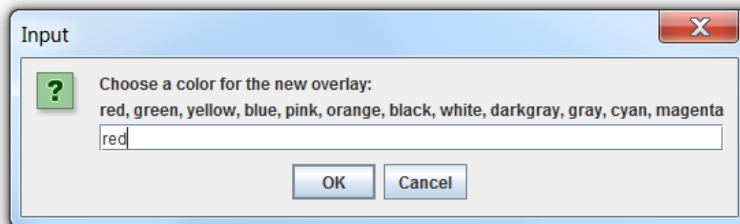


Figure 128. Select a color for the new overlay and wait for the display to update ...

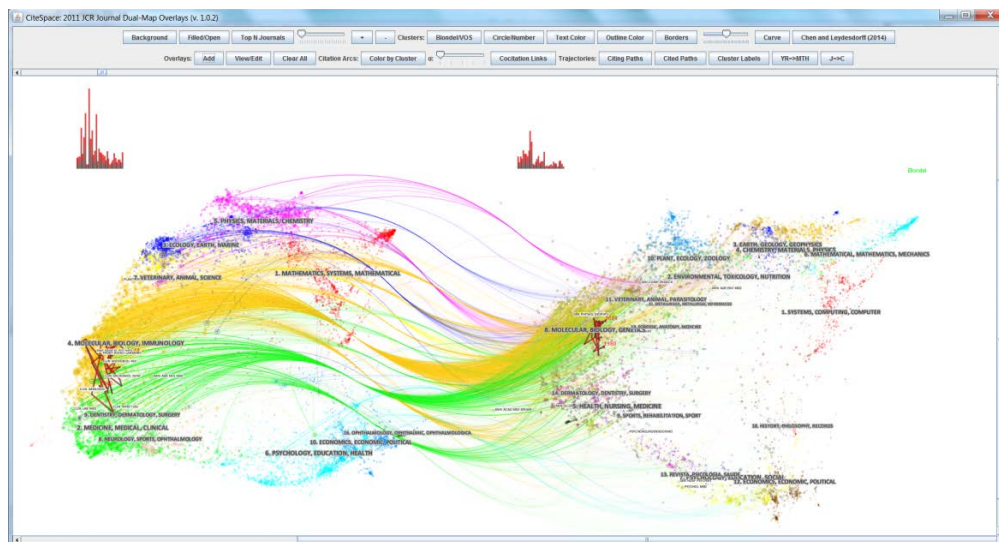


Figure 129. The result of adding a set of bibliographic records in the WoS format.

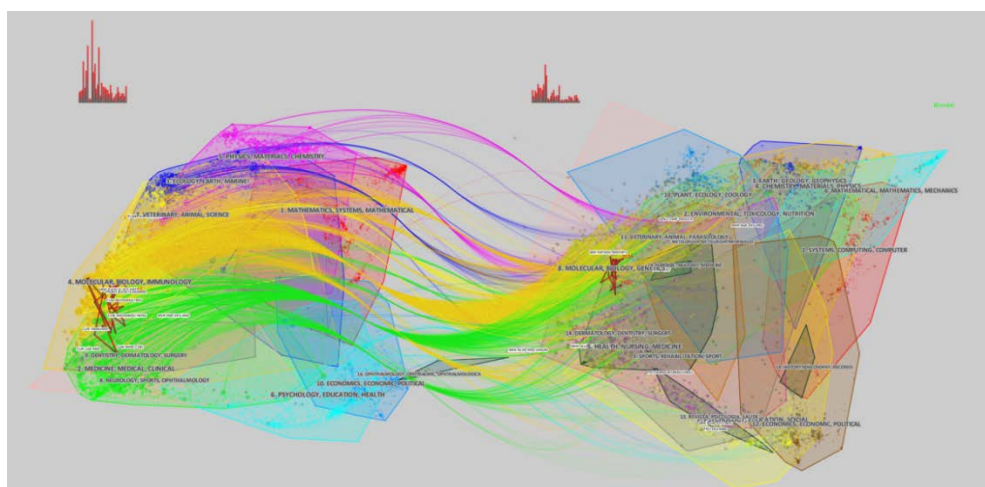


Figure 130. The Border button controls the display of the border of a cluster and whether to fill these areas in color.

## 8.4.2 Further Reading and Terms of Use

Technical details and several examples are provided in (C. Chen & Leydesdorff, 2014).

This function is made available for non-commercial and educational use. For commercial use, please contact me directly for further detail.

If you use the dual-map function in your publications, you should cite it as follows:

Chen, C., Leydesdorff, L. (2014) *Patterns of connections and movements in dual-map overlays: A new method of publication portfolio analysis*. *Journal of the American Society for Information Science and Technology*, 65(2), 334-351.

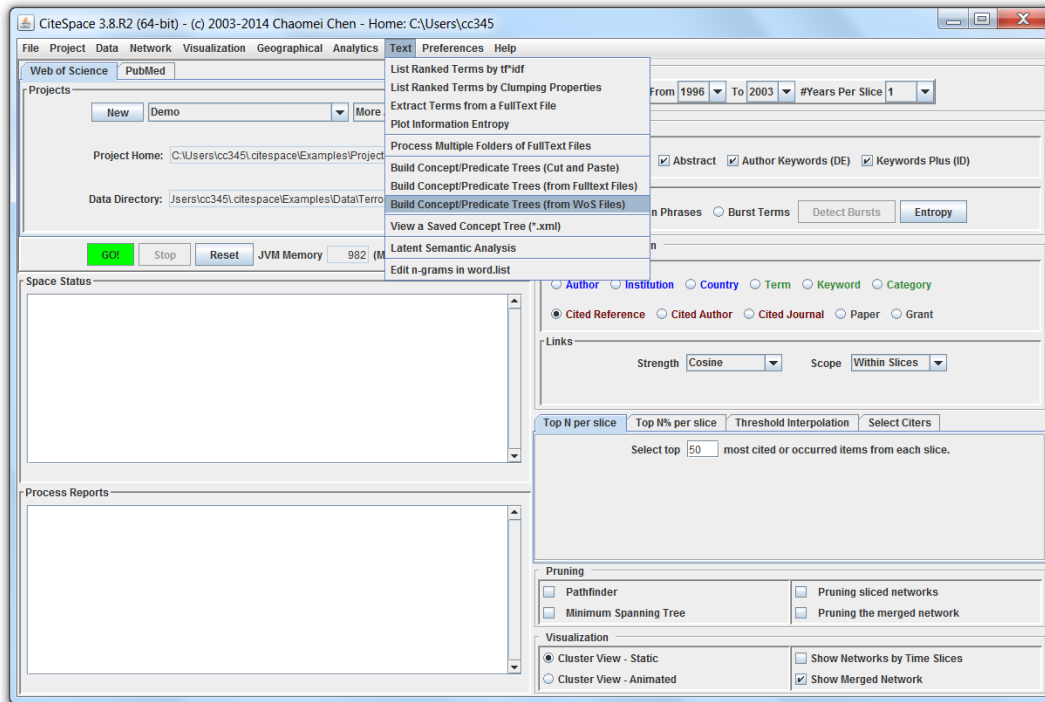
## 8.5 Menu: Text

### 8.5.1 Concept Trees and Predicate Trees

Concept trees and predicate trees in CiteSpace are generated from three types of unstructured text documents: 1) cut and paste text to an input window, 2) from full text files, and 3) from a folder of files in the WoS format, including the data files you downloaded directly from the WoS

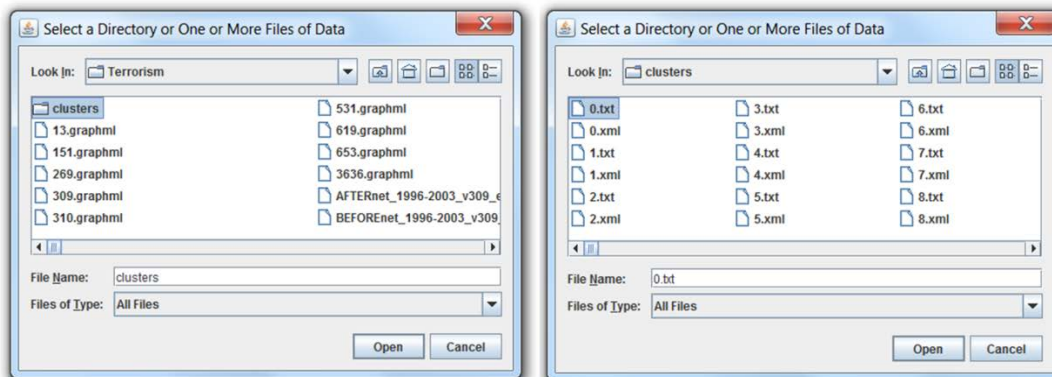
and intermediate files saved to the project folder after you performed the clustering algorithm to the current network.

The following example shows how to generate concept trees and predicate trees from the records that cited the largest cluster in the Demo project (i.e. the terrorism research). First, set the Demo project as the current project. Then follow the menu **Text ► Build Concept/Predicate Trees**.



**Figure 131.** Generate concept trees and predict trees.

You will need to select the file that represents the citing articles to the largest cluster of the Demo project. CiteSpace will show you a list of folders and files. Select the folder **clusters**, then 0.txt, which corresponds to cluster #0, the largest cluster.



**Figure 132.** Select the clusters folder of the Demo project and the largest cluster #0.

The concept tree window has three panels. The tree window shows a visualized concept tree. The context window shows the sentences that contain a concept, i.e. the node in the concept tree. The example below shows when you move the mouse cursor over the bioterrorism node in the Tree

window. Different phrases that contain the term bioterrorism are shown as the children nodes of the concept, for example, threat (of) bioterrorism, weapons and agent (of) bioterrorism.

The nodes near the top of the tree are major concepts and major concerns of the cluster. Thus we know that the largest cluster in the Demo project is really about bioterrorism, United States, biological attack, and effective response. These concepts, taken together, give us a fairly focused sense of the nature of the cluster.

To pane the visualized tree, hold down the left button of your mouse and move it around.

To zoom the visualized tree, hold down the right button of your mouse and move it up (zoom out) or down (zoom in).

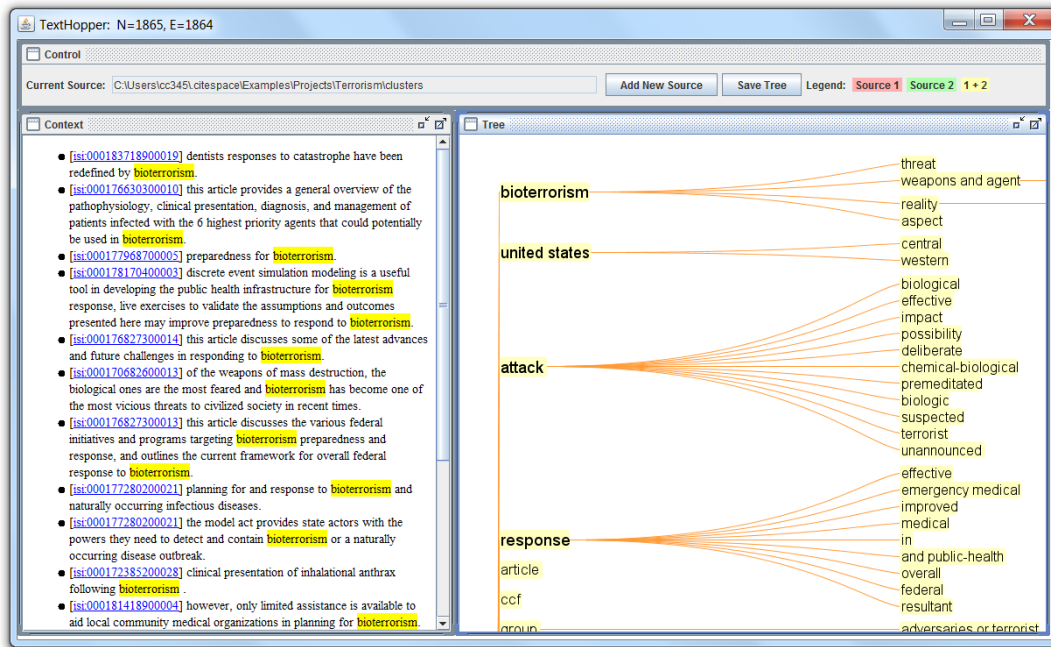


Figure 133. The concept tree of cluster #0 – bioterrorism in the Demo project.

In the Control window, you can add a new source to the existing concept tree. Here let's add the second largest cluster so that we can see what these two largest clusters have in common and where exactly they differ. Recall that the second largest cluster is labeled as PTSD – post traumatic stress disorder.

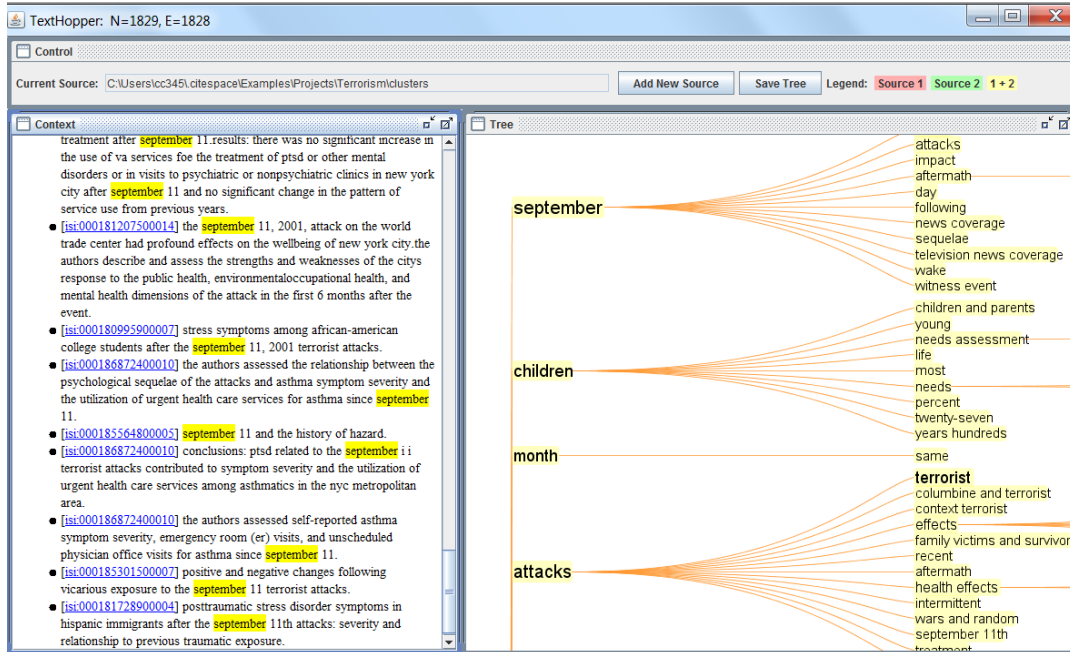


Figure 134. The PTSD cluster. Key concepts: September, children, same month, and terrorist attacks.

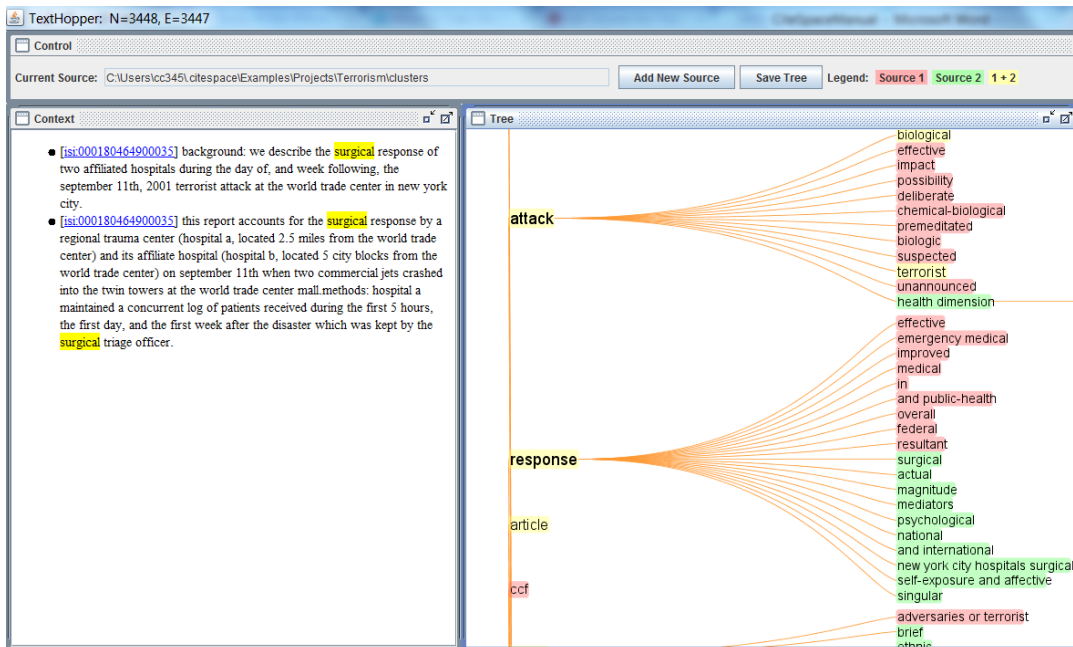


Figure 135. The concept tree of two sources. The bioterrorism cluster is in red. The PTSD cluster is in green. The overlap between the two is in yellow.

### 8.5.2 List Terms by Clumping Properties

Under the Text menu, you can find several functions dealing with text.

For example, Text ► List Ranked Terms by Clumping Properties, can sort terms by their clumping properties, i.e. how closely they tend to appear in text (Bookstein, Klein, & Raita, 1998). In the Demo project, the most prominent terms include terrorist attacks, world trade center, mass destruction, and biological terrorism.

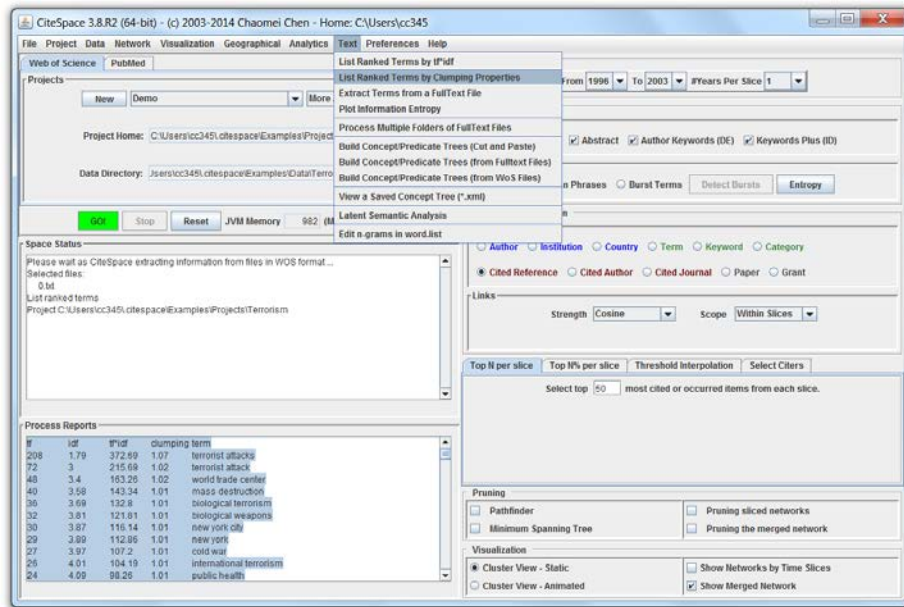


Figure 136. List ranked terms by clumping properties.

### 8.5.3 Latent Semantic Analysis

CiteSpace provides a somewhat underdeveloped Latent Semantic Analysis function under **Text** ► **Latent Semantic Analysis**. The Latent Semantic Analysis is based on a singular value decomposition of the term by document matrix. It is a dimension reduction method (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990).

Use the browse button to locate at least two data sources, i.e. folders of text files in plain full text or the WoS format. After select each data source, add it to the list using the button “Add to the List” then press the “Analyze” button. Then wait for it to finish ...

Once it is done, five most representative words in each dimension are shown in the user interface.

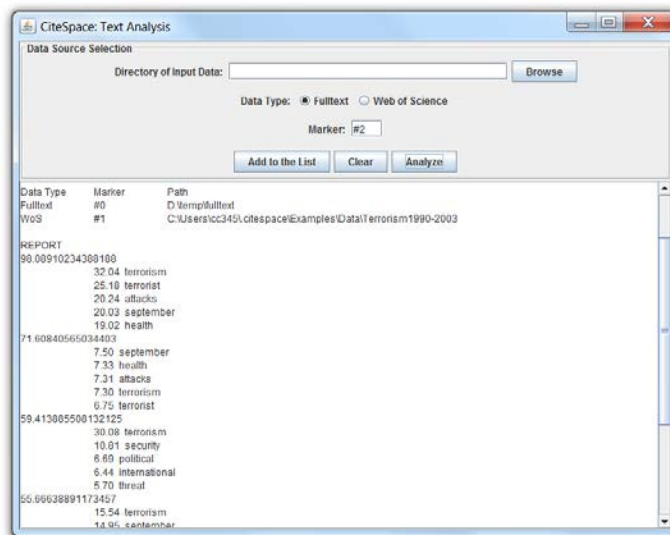


Figure 137. Latent Semantic Indexing.

Three coarse visualizations of the latent semantic space are provided for the three most prominent dimensions of the latent semantic space. Each visualization shows a mixture of terms and documents. You can zoom in and out, change the font size of labels, and the length of a label. That is about it. This function has been there for years, but it has not been actively developed.

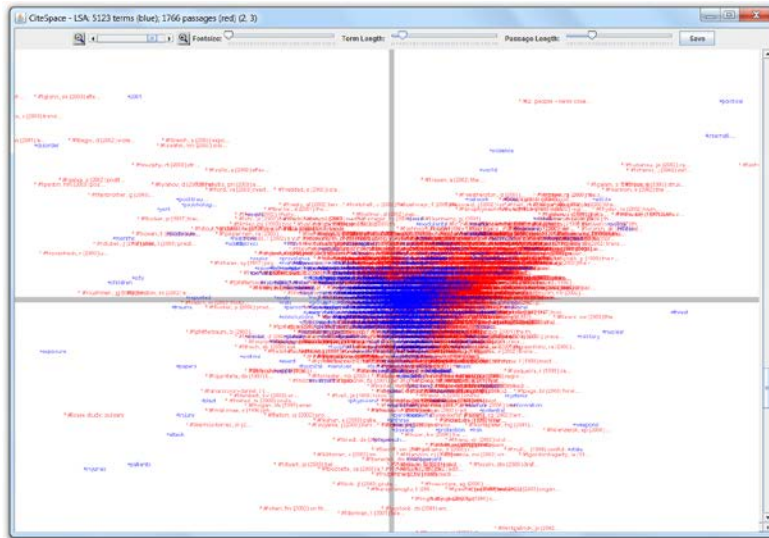


Figure 138. A visualization of the Latent Semantic Space, the 2<sup>nd</sup> and the 3<sup>rd</sup> dimensions.

## 9 Selected Examples

Here are some good examples I came across on the Internet. These examples are created by users with CiteSpace.

The following visualization is from [blog.sciencenet.cn](http://blog.sciencenet.cn) by Jie Li.

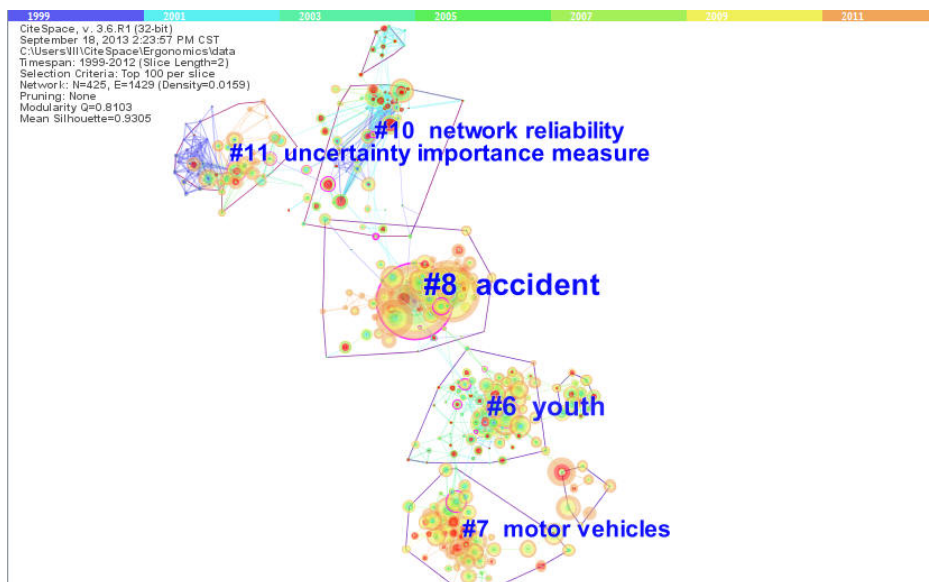


Figure 139. <http://blog.sciencenet.cn/blog-554179-729837.html>

<http://image.sciencenet.cn/album/201310/03/2133022vo2o5g7ppvgpp5i.jpg>



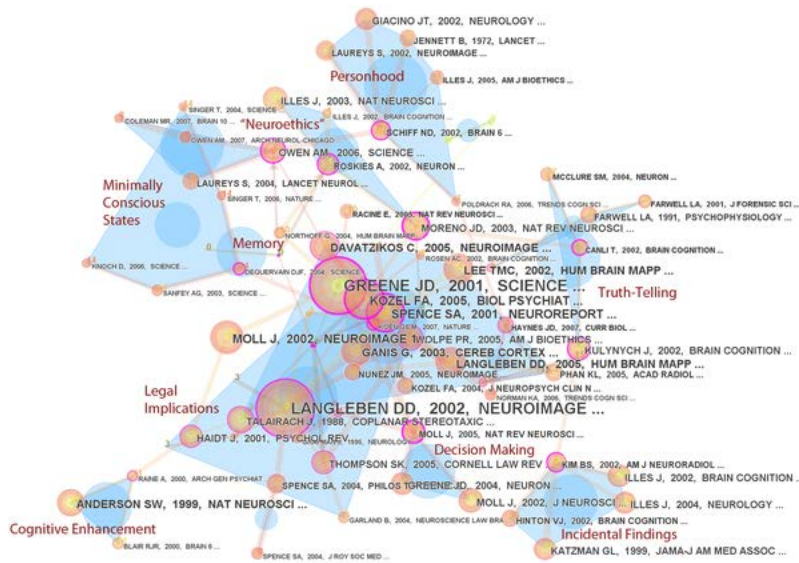


Figure 140. An article published in Plos One, Figure 3.  
<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0018537>

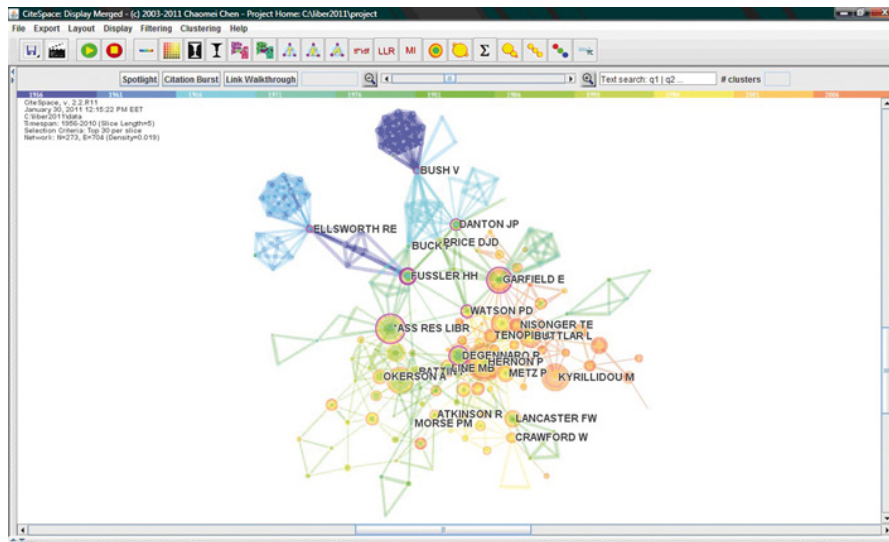


Figure 141. An article published in Liber Quarterly, Figure 4.  
<https://liber.library.uu.nl/index.php/lq/article/view/URN%3ANBN%3ANL%3AUI%3A10-1-113638/8398>

## 10 Metrics and Indicators

### 10.1 Information Theoretic

#### 10.1.1 Information Entropy

CiteSpace computes information entropy based on noun phrases extracted from records to represent the diversity of a data set.

### 10.2 Structural

#### 10.2.1 Betweenness Centrality

The betweenness centrality of a node in a network measures the extent to which the node is part of paths that connect an arbitrary pair of nodes in the network (Brandes, 2001; C. M. Chen, 2006; Freeman, 1977).

#### 10.2.2 Modularity

The modularity of a network measures the extent to which a network can be decomposed to multiple components, or modules. This metric provides a reference of the overall clarity of a given decomposition of the network (C. Chen et al., 2010).

The modularity change rate induced by a set of incoming information is considered to be a sign of a potentially important perturbation to a complex adaptive system (C. Chen, 2012).

#### 10.2.3 Silhouette

The silhouette value of a cluster measures the quality of a clustering configuration. Its value ranges between -1 and 1. The highest value represents a perfect solution. However, to ensure a sound interpretation in CiteSpace, it is recommended that you should balance the modularity and silhouette scores simultaneously (C. Chen et al., 2010).

### 10.3 Temporal

#### 10.3.1 Burstness

The burstness of the frequency of an entity over time indicates a specific duration in which an abrupt change of the frequency takes place (Kleinberg, 2002). In CiteSpace, citation burst and occurrence burst are both supported.

### 10.4 Combined

#### 10.4.1 Sigma

This indicator measures the combined strength of structural and temporal properties of a node, namely, its betweenness centrality and citation burst (C. Chen et al., 2009).

## 10.5 Cluster Labeling

### 10.5.1 Term Frequency by Inversed Document Frequency

### 10.5.2 Log-Likelihood Ratio

### 10.5.3 Mutual Information

## 11 References

- Bookstein, A., Klein, S. T., & Raita, T. (1998). Clumping properties of content-bearing words. *Journal of the American Society for Information Science*, 49(2), 102-114.
- Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2), 163-177.
- Carrot2: Open source framework for building search clustering engines. (2012). from <http://project.carrot2.org/>
- Chen, C. (2004). Searching for intellectual turning points: Progressive Knowledge Domain Visualization. *Proc. Natl. Acad. Sci. USA*, 101(Suppl.), 5303-5310.
- Chen, C. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology*, 63(3), 431-449. doi: 10.1002/asi.21694
- Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, 3(3), 191-209.
- Chen, C., Hu, Z., Liu, S., & Tseng, H. (2012). Emerging trends in regenerative medicine: A scientometric analysis in CiteSpace. *Expert Opinions on Biological Therapy*, 12(5), 593-608.
- Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. *Journal of the American Society for Information Science and Technology*, 61(7), 1386-1409. doi: 10.1002/asi.21309
- Chen, C., & Leydesdorff, L. (2014). Patterns of connections and movements in dual-map overlays: A new method of publication portfolio analysis. *Journal of the Association for Information Science and Technology*, 65(2), 334-351. doi: 10.1002/asi.22968
- Chen, C., & Morris, S. (2003, October 19-24, 2003). *Visualizing evolving networks: Minimum spanning trees versus Pathfinder networks*. Paper presented at the IEEE Symposium on Information Visualization, Seattle, Washington.
- Chen, C. M. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359-377. doi: 10.1002/asi.20317
- Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.
- Freeman, L. C. (1977). A set of measuring centrality based on betweenness. *Sociometry*, 40, 35-41.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14, 10-25.

- Kleinberg, J. (2002). *Bursty and hierarchical structure in streams*. Paper presented at the Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada. <http://www.cs.cornell.edu/home/kleinber/bhs.pdf>
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 265-269.
- Stasko, J., Gorg, C., & Liu, Z. (2008). Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2), 118-132.
- White, H. D., & Griffith, B. C. (1981). AUTHOR COCITATION - A LITERATURE MEASURE OF INTELLECTUAL STRUCTURE. *Journal of the American Society for Information Science*, 32(3), 163-171.