Network Graph Analysis and Visualization with Gephi

Chapter No. 4
"Creating a Gephi Dataset"
In this package, you will find:
A Biography of the author of the book
A preview chapter from the book, Chapter NO.4 "Creating a Gephi Dataset"
A synopsis of the book’s content
Information on where to buy this book

About the Author

Ken Cherven is a marketing analyst working in the automotive sector in Detroit, Michigan, USA. He has more than 15 years' experience working with proprietary tools from Microsoft, Cognos, Tableau, and Oracle, in addition to extensive experience using a variety of open source software applications including MySQL, SpagoBI, JasperServer, BIRT, Mondrian, R, Gephi, Exhibit, Omeka, and d3.

Ken also maintains the visual-baseball.com site, where he uses available open source and proprietary tools to analyze, report on, and visualize baseball information. The site features many of his baseball visualization projects, including a collection of more than 100 seasons of interactive pennant race charts.

One of Ken's current projects is to publish a visual history of major league baseball pennant races from 1901 through 2012, using a dashboard approach featuring horizon charts, box plots, bullet charts, and other visuals to tell the story of each and every race in a highly visual fashion. This book is scheduled for a 2013 release.

For More Information:
Network Graph Analysis and Visualization with Gephi

Network graphs have become an integral part of the visualization world, as users create examples that display connected networks across the worlds of social media, politics, corporations, travel patterns, and countless other themes. Many of these examples have been created using tools that require a significant time investment with a steep learning curve, making it a challenge for many potential creators.

Gephi helps to overcome these barriers by providing a powerful, yet easy-to-use framework that allows users to spend more time on creating and deploying network graphs, while spending far less time coding. If you have interesting datasets that will provide the foundation for compelling graphs, Gephi will help you to quickly create, customize, and deploy your graphs to a wider audience.

The goal of this book is to help as many people as possible to learn the basics of network visualization through Gephi, and to empower each reader to create his own unique visualizations that can be shared with a wider audience.

What This Book Covers

Chapter 1, Installing Gephi, will teach you how to quickly and easily install and configure Gephi.

Chapter 2, Creating Simple Network Graphs, will teach you how to use the default Gephi settings to quickly create your own network graphs.

Chapter 3, Exploring Additional Layout Options, will show you the multiple ways to create and view network graphs. You will learn how to use several available layout options in Gephi to make your own compelling visualizations.

Chapter 4, Creating a Gephi Dataset, will help you create your own Gephi datasets using both the Gephi data laboratory, as well as spreadsheet software and the MySQL database.

Chapter 5, Exploring Plugins, will show you a number of plugins that the Gephi community offers to expand the core capabilities of the software. In this chapter, we'll learn how to install and configure a few of the best examples.

For More Information:
Chapter 6, *Advanced Features*, will take you beyond the basics to begin exploring features such as filtering, querying, and ranking to help you create powerful and informative network graphs.

Chapter 7, *Deploying Gephi Visualizations*, helps you go further than just creating your own network graphs. You will learn how to export them into other formats or to the web for others to view.

Appendix, *Network Visualization Resources*, is a resource to help expand your capabilities, with helpful links to Gephi resources and visualization websites, books, and software.

For More Information:  
Creating a Gephi Dataset

So far we have worked with existing datasets as we created our network graphs, and now it's time to introduce you to the process of creating your own data. There is more than one way to do this, so we'll begin with the simplest approach, and then progress through some more advanced options.

By the end of this chapter, you will be familiar with the following methods for creating or importing data in Gephi:

- Inputting the data manually using the Data Laboratory tab in Gephi
- Using spreadsheet software, such as Microsoft Excel or OpenOffice Calc
- Finally, working with datasets from a MySQL database

You will also gain a greater understanding of how to define nodes and edges in your dataset, which will speed your ability to create network graphs.

So let's start by making sure we're absolutely clear on the basic requirements for creating a dataset.

Basic data requirements

At the most fundamental level, there are just two data components needed to create a network graph in Gephi.

- **Nodes**: These form the foundation for any network graph, as they represent all of the entities within the data. There is a single node for each unique entity in our data.
- **Edges**: These are the connections between entities, defining the sort of relationship that exists between nodes. There are typically many more edges than nodes in a normal graph dataset, because edges represent each and every connection between nodes.

For More Information:
Assume an example where our datafile contains 200 colleges and universities across the world. Each of these institutions has worked with others to produce scientific research. We would like to answer several questions using this data, including the following:

- How much cooperation is there between academic institutions?
- Do certain institutions produce more research than others?
- Is the research clustered within smaller groups, or do many universities work with a lot of partners?
- Are there specific topic areas dominated by a few universities?

To answer these questions, we will need several data elements. We will require an entry for every university or college that has published scientific research. Each of these institutions will then have a single node in our datafile. What about those institutions that produce far more research than others? Shouldn't they have multiple rows in our dataset? Good question, but the answer is no. We already noted that each entity will be represented by a single node, but there are a couple of ways we can address this issue in order to produce an informative and accurate graph.

**Sizing nodes and edges**

We have several data options at our disposal to make our graph tell a compelling yet accurate story. The following paragraphs contain a few suggestions on how we might wish to achieve this goal. If you wish to gain a deeper understanding of how nodes and edges are used before moving on, I suggest that you download and read the Graphs chapter from Easley and Kleinberg that was recommended in Chapter 1, *Installing Gephi*, of this book.

First, we could provide a size element for our node data that corresponds to the number of projects published by that institution. So, if Harvard was involved in 39 projects and Yale had a hand in 14, we could reflect this in our data by using a size element, and letting Gephi know which field to use when creating and displaying nodes. We can do this using the edit node properties option. Just be careful not to exaggerate the size ratio; many graphics programs (Gephi included) set the radius or diameter of a circle, rather than the area, which provides the true measurement of difference.

A second option is to refrain from sizing the nodes, and simply use the number of edges to demonstrate the influence of one institution versus another. Since every research project will have edges connecting the multiple publishers of the research, universities involved in many efforts will have a greater number of edges flowing into and out of their node.

For More Information:
Or, we could employ both methods in tandem, which would then show us both the proportion of projects worked on by one college as well as with whom they worked. This approach would allow viewers to get a good glimpse into the overall magnitude of research projects by school, as well as the number of connections to partner collaborators.

An additional option is to use color to differentiate between values. In this example, the country of the academic institution provides an ideal opportunity for working with color. The use of specific colors could then help us to understand academic cooperation across international borders.

There is one more option we could employ, one which involves applying sizes to the edges. This is commonly known as weight, which is an indication of the intensity or strength of the connections between nodes. For example, if Harvard and Yale worked on five projects together, we might want their edge weight to be five times greater than for institutions that worked on only a single project together. This is a highly effective way to show the magnitude of relationships, but must be used carefully so as not to obscure the overall impact of the graph. Edge weights can also be utilized within certain layouts and statistical measures to add to our understanding of the data.

So, you see that although the concept of nodes and edges is fairly simple, we have the ability in Gephi to glean multiple insights based on how we structure our data. We can effectively make our data to be more intelligent before we even begin creating a graph. Plus, it is far easier to build these capabilities into the base dataset than it is to make modifications once the data has been loaded into Gephi.

Now that we know what to build into our data, let's begin by creating a dataset using the Gephi Data Laboratory.

Building a datafile in Gephi

The simplest method to create a basic Gephi datafile is to use the built-in Data Laboratory, which gives you the ability to define nodes and edges manually. Since it is such a manual process, I can't recommend using it for anything beyond a small datafile. Still, it's a nice option to have, and it can also be used to edit or append records to data imported from outside of Gephi.

If you have your copy of Gephi open, go to File | New Project, and then navigate to the Data Laboratory tab (hint: it's next to Overview at the top of your screen). What you should see next is a very empty window, devoid of anything beyond a small menu options at the top and bottom of the workspace. We're going to begin filling that space by manually adding some nodes and edges.
Adding nodes
Click on the Add Node button, and you should see a dialog box requesting a label for your first node. We’ll give it the highly creative Node_1 label, and click the OK button. You will now see a Node_1 entry, with Node_1 serving as both the label and node identifiers. Meanwhile, Gephi has created an ID value for this node, independent of anything we have done.

Let's add five more nodes so we will have enough to make a reasonably interesting graph. Follow the same steps we just used for Node_1, and create nodes 2, 3, 4, 5, and 6. When you are done, we can move on to creating some edges.

Finished? Good. I hope you started to get the feeling that creating nodes manually would not be the smart way to go if we had three or four hundred entities we need to graph. If you're anything like me, it would quickly become an incredibly tedious process. In any event, we were able to create our modest set of nodes, so let's move on to add some edges.

Adding edges
Adding edges is similar to creating nodes, with one major exception—edges are generally not independent. An edge will typically connect two nodes only once, so the Gephi process for adding edges gives us a little bit more assistance. When we click on the Add Edge button, the following dialog screen pops up:

Note that we are given a couple of options here, starting with choosing a Directed versus Undirected connection. A directed edge is a one-way connection between two nodes, while an undirected edge implies no directional relationship. In a sense, it represents a two-way path between the connected nodes.

For More Information:
Next, we specify the **Source node** and **Target node** values, with the understanding that this distinction is meaningful only when we have a directed relationship. Let's connect Node_1 and Node_2, using an undirected edge.

You'll notice Gephi has created an entry in the Data Laboratory showing this edge as an undirected type. Let's create a few others—make your own edges, and we'll reconvene in a moment.

Ready? You may have noticed that when you create an edge between two nodes, Gephi automatically updates the drop-down list so you don't inadvertently create duplicate edges. There are cases where two nodes might have parallel edges, but this feature is not currently supported in Gephi. Recall that we can apply edge weights to show stronger connections, and we'll look at that in a moment. For now, let's see what our graph looks like based on the edges we just created. Here's what I got—remember that yours will likely be a bit different, unless you created the exact same connections:

![Graph Image]

I dressed mine up just a bit by changing the node color and adjusting the edge thickness with the slider control, but you should have something that vaguely resembles this graph. Congratulations—you've just created your first network graph using your own data!

Next, we'll move on to the more practical option of sourcing our data from a spreadsheet program, such as OpenOffice Calc or Microsoft Excel.

For More Information:
Creating a Gephi Dataset

Using spreadsheet files in Gephi

If you're planning to work with datasets encompassing more than a handful of nodes and edges, I strongly encourage you to use OpenOffice Calc, Microsoft Excel, Google Spreadsheet, or Zoho Sheet to do the primary work. You can then easily read the data into Gephi as a .csv file, using any of four delimited formats—comma, semicolon, tab, and space.

Creating and importing a spreadsheet

Here are the steps you'll need to take for this approach:

1. Create a nodes file in your favorite spreadsheet software, using the following fields. Be sure to include column headers, because it will make the import process easier to follow:
   - Nodes in the form of a brief identifier or abbreviation
   - Id as a unique numeric identifier
   - Label as a more descriptive name for the node

2. Create an edges file using the following headings:
   - Source refers to the originating node for an edge.
   - Target refers to the target node at the other end of the connection.
   - Type can be directed or undirected, depending on whether your data is nondirectional or directional.
   - ID can be provided, or Gephi will create it automatically.
   - Label can describe the connection in more detail (optional).
   - Weight should be used to show the frequency of connections between two nodes, assuming you wish to display that in your graph; if left blank, Gephi will provide an equal default value for every edge. Decimal and integer values can both be used to define the edge weights.

3. Save your files to a .csv format.
4. Go to the Gephi Data Laboratory tab, and select the Import Spreadsheet button.
5. Import the nodes file.
6. Import the edges file.

That's all there is to get your spreadsheet data into Gephi and begin making some beautiful visualizations.

To illustrate this process using real data, we'll work with a couple of files I previously used to create a network graph in Gephi. Ready? Let's begin.

For More Information:
Importing spreadsheet files

The two files we’ll be working with use baseball data that looks at the number of trades between teams over a 110-year period. This is a nice dataset with which to work, because it has just a few dozen nodes, but also a lot of variation in the edges. Some teams have rarely traded with one another, so the edge weight of their connection will be minimal, while others will have much thicker lines, indicative of frequent transactions. You can download these files at https://app.box.com/s/w6yfjkp8j0kpopp94ui6.

Study the files for a moment so you become acquainted with the stories the data might tell. This may also help guide you toward a particular graphing approach, although I strongly recommend that you test several methods once the data has been loaded. If you’re like me, what seems good initially doesn’t always pan out. Fortunately, we have a wide range of options for creating our graph.

Alright, let’s get started. First, we’ll open the nodes file. Click on the Import spreadsheet button and find the node file you just downloaded. You should see something like this on your screen:

![Import spreadsheet dialog box]

---

For More Information:
Make sure the **Nodes table** option is selected from the **As table** drop-down list. Notice how the **Preview** section displays the **Nodes**, **ID**, and **Label** fields, showing the first several rows of the data. In this case, the **Nodes** and **Labels** house identical values. We could make the labels more explicit in the Data Laboratory once the data has been imported, but for now let's leave them unchanged. Then select **Next**, and verify all settings on this screen:

Take note of the final checkbox, which defaults to forcing nodes to be created as new ones; if you are merely updating existing nodes, then uncheck this box.

---

**For More Information:**

So now our nodes are set, and we'll need to import the edges that connect them. We'll follow a very similar procedure, albeit with slightly different options. Click on the **Import spreadsheet** button once more, and choose the **Edges table** option from the drop-down list. Find your file, and you should see something like this:

Now we have more fields to work with, as we discussed earlier in this chapter. We have both **Source** and **Target** fields, **Type**, **ID**, **Label**, and **Weight**. Notice that the **ID** and **Label** fields are both not populated. Gephi will automatically create an ID value, so we needn't worry about that when we create our source file. Labels are not always used for edges, although if you have specific cases where you wish to see them, by all means create those values either in your spreadsheet file or via the Gephi Data Laboratory.

Click on **Next**, and you'll be able to see the settings for the edge fields, much as we did for nodes a moment ago.

---

For More Information:  
Creating a Gephi Dataset

At the bottom of the screen, you will see an option to create missing nodes, which is checked by default. In the event that our nodes file failed to contain every value represented in the edges file, Gephi will create the missing values for us.

We now have our data imported, and can move on to create graphs using this data.

Importing MySQL data

MySQL is a very popular database for users of open source software, so it feels like a perfect choice for Gephi users to download data from their MySQL-based applications.

We'll do a simple introduction to importing data from a MySQL database to load into Gephi, using the same data we just saw for the spreadsheet example. For database data, the process begins by selecting File | Import Database from the Gephi menu, which will present you a list of options to be filled in regarding your database. After you populate the various fields, you should wind up with something along these lines:

For More Information:
You'll need to populate the following fields:

- **Configuration Name**: This can be any name that makes sense to you for your project.
- **Driver**: For this field select MySQL. Gephi also works with SQLServer, Teradata, PostgreSQL, and SQLite, but we'll focus on MySQL here.
- **Host**: This will be the domain where your MySQL server is located.
- **Port**: This is typically **3306** for MySQL.
- **Database**: Here select the database where you have your node and edge tables.
- **User Name**: This will bear the ID you use to access the database.
- **Password**: This is your user password for the database.
- **Node Query**: This field allows you to select a subset of data using a *WHERE* clause, or you can simply select all records, as shown in the preceding screenshot.

For More Information:
Creating a Gephi Dataset

- **Edge Query**: This field allows you to select all edges or a subset; make sure your node and edge queries are consistent!

You can also test your connection before proceeding with the queries by clicking on the **Test Connection** button. That's all you'll need to get started using Gephi with MySQL. If you wish to use one of the alternative databases, consult the Gephi wiki and forum for further information, or take a look at the documentation for your specific database type.

Once you have connected to your database, you should find the process to be very similar to what we walked through in the spreadsheet example. Click on the **OK** button to continue. You will now see the **Import report** window, verifying your data structure. After reviewing the settings, click on **OK** again to create your new graph.

**Saving your file**

Regardless of how your data was created, when you save a file, Gephi will store all the layout and attribute information in a file with the `.gephi` extension. To save a file, simply use **File | Save**, or press **Ctrl + S**.

**Summary**

In this chapter you learned how to create and import Gephi data. Specifically, you should now understand the concepts detailed below.

Nodes represent the items or entities within a network graph, and edges are the lines that connect nodes to one another. Each node or edge can be sized to represent the relative weight of both an item and its connections to other nodes in the network accurately. Gephi makes it quite simple to specify these weights using the Data Laboratory.

You have also learned how to import data from a spreadsheet file into Gephi. This approach will enable you to work with larger files and specify node sizes and edge weights without having to use the Data Laboratory.

Finally, we learned how to use a MySQL database connection to create files to be imported into Gephi. This can be an excellent option if you have data that is already in the MySQL format, and it is often the best choice for working with large datasets.

Now that you are comfortable creating and importing data using Gephi, it's time to explore some of the available plugins that will make it easy to turn your data into meaningful graphs and insights.

---

**For More Information:**

Where to buy this book


Alternatively, you can buy the book from Amazon, BN.com, Computer Manuals and most internet book retailers.