

Analyzing and visualizing your Twitter networks in Gephi

Rense Corten, Bas Hofstra & Joris Broere, Department of Sociology, Utrecht University

Introduction

This instruction consists of three parts:

1. We will first get familiar with analyzing and visualizing social networks in Gephi, by exploring a small and easy-to-manage dataset of high school friendships.
2. Next, we will walk you through downloading network data from Twitter.
3. Finally, we will repeat part 1, but this time on your freshly downloaded Twitter dataset, which probably will be much larger than the practice dataset.

We assume that you have prepared the following:

- You have a Twitter account
- You have installed R (and, preferably, R-Studio)
- You have installed Gephi

Let's get started!

Part 1: first steps in Gephi with a toy dataset

1. Go to <https://www.uu.nl/en/events/clue-training-6-gephi-network-analysis-and-visualization-software> and download the data: knecht08b_actors.txt and knecht08b_ties.txt.
2. Start Gephi
3. From the dialog that appears after startup, or from the file menu, start a new project
4. Go to "Data laboratory"
5. Choose "import spreadsheet"
6. Choose knecht08b_actors.txt
7. Set the right separator (tab), and set as "nodes table"

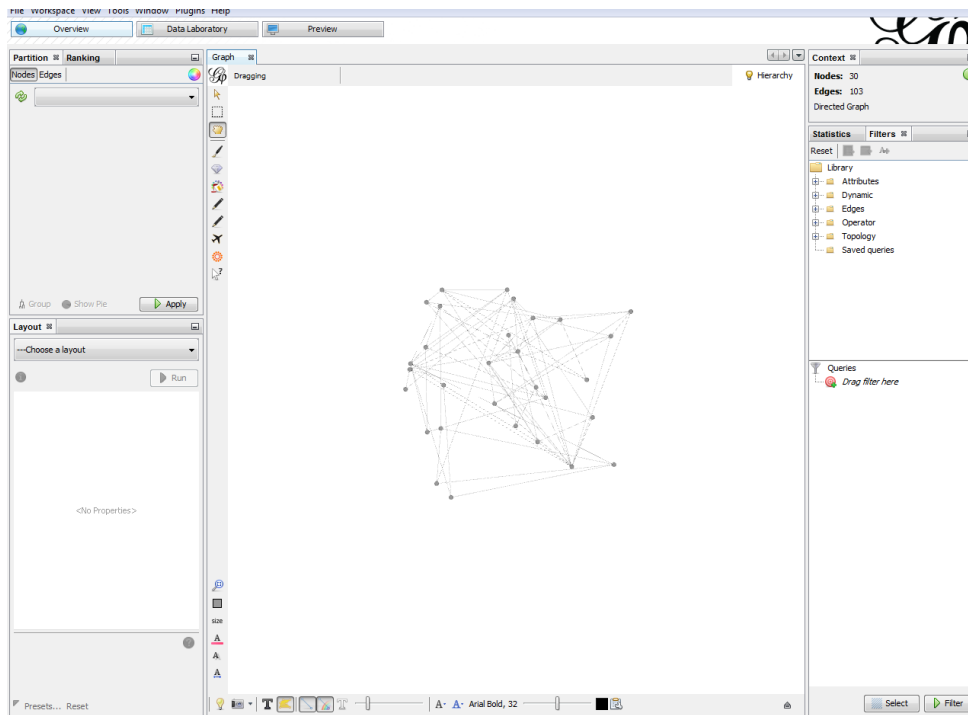
8. Click next → you should see a neat spreadsheet now:

Nodes	id	Label	name	sex	achomed	landno
1	1		1	female	sometimes true, sometimes not	4
2	2		2	male	true	1
3	3		3	male	sometimes true, sometimes not	1
4	4		4	female	true	1
5	5		5	male	true	1
6	6		6	male	true	1
7	7		7	female	true	1
8	8		8	male	sometimes true, sometimes not	1
9	9		9	female	true	1
10	10		10	male	sometimes true, sometimes not	1
11	11		11	female	sometimes true, sometimes not	1
12	12		12	female	sometimes true, sometimes not	1
13	13		13	male	true	1
14	14		14	male	very true	1
15	15		15	female	true	1
16	16		16	male	not true	1
17	17		17	female	very true	1
18	18		18	male	true	1
19	19		19	male	true	1
20	20		20	male	true	7
21	21		21	male	sometimes true, sometimes not	1
22	22		22	male	sometimes true, sometimes not	1
23	23		23	male	sometimes true, sometimes not	7
24	24		24	female	very true	1
25	25		25	male	sometimes true, sometimes not	1
26	26		26	male	sometimes true, sometimes not	1
27	27		27	female	true	1
28	28		28	male	true	1
29	29		29	female	sometimes true, sometimes not	1
30	30		30	male	very true	1


9. Click “import spreadsheet” again

10. Repeat the process for the file with ties, but this time choose “edges table” as the table type. Click “next”, then “finish”.

11. Go to the “Overview” tab. You should now see your network.




12. Play around with the various layout procedures (click “Run” to apply). Try at least “Force Atlas”, “Force Atlas 2”, “Fruchterman Reingold” and “expansion”. Notice that every layout algorithm can be tweaked by changing parameters.

13. Now add actor attributes. In the upper left corner, under “appearance”, click Nodes, then “attributes”, and then the  button. You can now select the actor attributes from the nodes file. Select “sexd”, and then apply. Next, have a look at “acthomed”, or the respondents’ answer to the statement “I always do my homework”. Do you see any sign of clustering here, based on one of these attributes?

14. No go back to the “overview” tab. At the right side of the screen, click “statistics” and then “run” next to “average degree”. Gephi shows you the degree distribution(s!), but also adds the results to the dataset in the “data laboratory”.

15. Also run the “network diameter” procedure, which gives you various centrality measures based on network diameter.

16. Go back to the data laboratory. Notice how the individual-level data now have network measures added. You could, from here, export the data (“export table”) and import them into your favorite data analysis software to further analyze the data. For example, is there any association between popularity (indegree) and doing your homework? (What would be your hypothesis?)

17. We can now make degree visible in the network map. Go back to “overview”. In the appearance tab, click the  (size) button and “attribute”. In the drop-down menu, choose indegree. Set “min size” to 2 and “max size” to 15, and click “apply”. The nodes are now sized proportional to their popularity.

18. Finally, for a pretty publication-ready picture, we can export the network map. Go to the “preview” tab, and click “refresh”. Try some of the presets (upper right). The network map can be exported in various formats (including vector graphics) by clicking “export” in the lower left corner.
19. For more details on visualization, try one of the tutorials at <https://gephi.org/users/quick-start/> and <https://gephi.org/users/tutorial-visualization/>

Part 2: Downloading Twitter data

1. Make a Twitter account:

Note: if you already have a Twitter-account *and* have added your phone number to this account, you can skip this step. If you do not want to use your regular Twitter-account for this computer lab, you can make a new Twitter-account, where you’ll need to sign up with another e-mail address than the one you use for your regular Tweets. Follow the steps below:

- a. Go to www.twitter.com
- b. [Make a new Twitter-account](#)
- c. In order to be able to download Twitter-data you must declare a telephone number. Please do so by going to settings → your Twitter data → add phone number. Wait for the verification code and add the phone number to your account. (note: sometimes there are problems with leading zeros, e.g., +31624587121 will work, but +310624587121 will **not** work).
- d. Congratulations, you’re now officially a member of the Twitter community which is over 500 million strong, and more importantly; this is the first step in downloading Twitter-data

2. Make a Twitter-app:

One way to mine Twitter-data is via the statistical program R through the Twitter-rest-API. This is also the method we are going to use in this computer lab (for other methods, see e.g., Ackland 2013:73). Therefore, you need to make a Twitter-app! This sounds more difficult than it is, we will guide you through making this app via the steps below:

- a. Go to <https://apps.twitter.com/>
- b. Sign in with your (newly created) Twitter-account
- c. If you have never developed an application for Twitter, the new screen should say “You don’t currently have any Twitter apps”.
- d. Press “Create New App”
- e. In the new screen you have to fill out a number of fields, you can for instance name your new app “ISMN Twitter mining”, or whatever name you deem necessary. Also fill out the description as you find suitable, for instance some information about the course and why you need to make this app. The website can be *any* website that works, for instance you linkedin/Facebook/personal website; as long as the URL works (include “http://www.” in the link). Agree to the terms and conditions and create the app.
- f. Congratulations, you have created a Twitter application!

3. Configure R

We will use R to handle the download of data. If you’re not familiar with R, see <https://www.r-project.org/> for more information. We also find the R-Studio interface convenient, which is a

convenient visual representation of R, e.g., you can see which packages are loaded and which data sets are declared. R-studio also works slightly better with point and click). Download the Twitter-mining R-script from <https://www.uu.nl/en/events/clue-training-6-gephi-network-analysis-and-visualization-software> in a folder of your choice (e.g., designated course folder).

- a. Open R on your computer (it is important that you do *not* run the 64-bit version; this will not work with the standard 32-bit Java that most Windows computers are equipped with). Better is to run the syntax from R-studio, which is nicer visual representation of the R-program.
- b. R is an open source program that works with *packages*, this means that R-users can build new functions and add these to the libraries in R. Users that are interested in such packages can in turn download and add these packages to R. We are definitely interested in a few packages, since they help us download Twitter-data.
- c. First we are going to need to install a few packages and load them into R. You can do this by following the “### C ###” steps in the R-script that you’ve downloaded.
- d. We now have created a situation in which we can actually start to download twitterR data through the “twitterR” package.

4. Download selections of Twitter-data

We are now going to download Twitter-data through the package twitterR. There are many more ways to get data through this package, you can always visit the help page of this package here: <https://cran.r-project.org/web/packages/twitterR/twitterR.pdf>

- a. You must now gain access to the Twitter-API via your R-console.
- b. Go to: <https://apps.twitter.com/>
- c. Select the application that you’ve created under step 2
- d. Click the tab “Keys and access tokens”
- e. Click “Create access token”
- f. Now you can see your Consumer Key (API Key), Consumer Secret (API Secret), Access Token and Access Token Secret. These need to be copied into your R-script: please follow the “### F ###” steps in the R-script.
- g. Now that we are able to download Twitter-data, we need to first explore some functions of the twitterR package that are useful
- h. We can download mentions of specific words, which is done with the function “xxxx <- searchTwitter(“lookupword”, n = xxxx, retryOnRateLimit = xxxx)”.
- i. In this way, we can download Tweets containing specific words, but we will also know who made the original Tweet, who retweeted the Tweet, the screennames of users, sometimes the geo-location from where the Tweet is done, the platform (e.g., iPhone) from where the Tweet was done, and some additional information.
- j. For this tutorial we are interested in information surrounding the hashtag “#somethingyouwanthere”. (note: change to something you find interesting.)
- k. Go back to the R-script and follow the steps under “### K ###”.
- l. Now that we have created a dataset in excel format, open this dataset in SPSS or the statistical software of your choice (first, open the data in excel and give the first column a name and save the excel-file again, or else SPSS won’t open the data).
- m. **Question: How many Tweets have you downloaded?**
- n. **Question: How many retweets on average does a Tweet get?**
- o. **Question: How many *original* Tweets have you downloaded? (tip: select only if *isretweeted* is FALSE)**

5. Creating network data from the downloaded Twitter

Now we want to see whether we can obtain a network from this dataset. Consider that on average Tweets are Retweeted more than five times. You might see this as directed relationships; those who Retweet an original Tweet have a tie with the original Tweeter. Also, those who reply on a Tweet have relationships to another Tweeter, we do not take these into account for this tutorial. So what we want to do now is to get a dataset that shows the directed Retweet network of the hashtag “#somethinyouwanthere”. Therefore, we need to do some cleaning on the data in R, i.e., to obtain screennames from the Tweets (we already have the original and Retweet screennames).

- a. Go back to the R-script “ISMN_twittermining_In08.r” and follow the steps in “### A ###”.
- b. Now we have created an actorlist and an edgelist. Open each file with a text editor and have a look. As often with social network data, the information is separated into two files:
 - a. **yourfilename_actors.txt** contains information about the actors (in this case, high school students), such that each row in the data refers to a person.
 - b. **yourfilename_ties.txt** contains information on relations between the students: each row in the data refers to a tie from one student to another. The numbers for the students are identical to those in the column “Id” in yourfilename_actors.txt.
- c. **Question: how many actors and how many ties are in the data?**

Part 3: Visualizing the Twitter network

Repeat the steps of part 1 with your own twitter data!