# Social Network Analysis – Winter 2022
# Lab 1a: Descriptive Network Analysis – Collecting and Visualizing Data

## Deadline: Friday, January 29th at 11:59 pm

The purpose of this lab is to learn how to conduct descriptive network analysis using the statistical programming language R. Using the "vosonSML" package, this assignment will make use of a network data set you collect from one of three social media platforms (i.e., Twitter, YouTube, or Reddit) by defining either hashtags (e.g., #MeTooMovement) or urls (e.g., a metoo movement thread on Reddit YouTube). A network is generated from the interactions (e.g., @mention and co-commenting) between users/actors included in the same hashtags or discussion thread. For example, Twitter user A replies to/retweets a tweet by B. It creates a link from A to B. On Reddit or YouTube, user X starts a thread and Y comments on the thread. Then, the interaction creates a link from Y to X. You will be visualizing and interpreting individual and global network properties of this type of network.

We will cover this material in lab section on **January 14th**. However, feel free to work ahead, and reach out to Brennan if you have any other questions. Thanks!

**General Instructions:**

1. This lab has two parts indicated with Roman numerals (**I, II**) in the outline below.
2. Prepare a report that includes your responses to **all the questions** for all four parts outlined below. Label your responses with the instruction and prompt number (for example, "9"). Incorrectly labeled responses may receive a lower grade. For each response in your report, you should report your results and interpret them as specified in the prompt. Insert network images into your report in the appropriate places. In RStudio, you can click "Export / Copy to Clipboard" and paste directly into the Word document. You will be graded primarily on the completeness and accuracy of your responses, but the clarity of the prepared report will also affect your grade. While students may work together to perform the analysis, each student must execute his or her own code, and is responsible for writing the narrative in the report and submitting it.
3. Upload your report as a PDF, R code script and RData file to the Lab 1 Assignment in Canvas by **Friday October 15th, at 11:59 pm.**
4. Please delete the instructions from your final hand-in.
5. You should install R, verify your installation, and collect your data as soon as possible so that you can receive technical assistance from the TA if needed. DO NOT WAIT UNTIL THE LAST MINUTE TO START THE LAB.
6. Link to the R project and relevant downloads: https://www.r-project.org/
7. After installing R, please install R Studio for running R: https://www.rstudio.com/ All software is free to download and works on both MAC and PC.

## PART I: Network Data Collection from Social Media (20 points)

For this lab, you will collect data from one of three social media platforms (i.e., Twitter, YouTube, and Reddit), save data from the search, create networks from the data, and compare the differences among networks.

**Part I Data Collection Instructions:**

1. Choose a topic for your hashtags or urls:
   You can decide hashtags or urls based on personal interests, research interests, or popular topical areas, among others. You have flexibility in selecting your list. For example, you can search for commercial brands, celebrities, countries, universities, etc. It will be most useful if you choose a topic that is seemingly controversial. A controversial topic means that at least two different opposing stances exist to discuss a topic. For example, #MeToo movement includes people who completely support vs. others who have concerns (e.g., a Vox article). Because you might want to see some separated communities in your communication detection analysis, think about a topic that might have interesting conversations among people who have different patterns of communication on social media.

2. If you choose to collect data from either Twitter or YouTube, request an API key:
   With vosonSML, you can collect network and text data from, Twitter, YouTube and Reddit. However, for these data sources you will require access to the respective application programming interfaces (APIs). This section provides some information on how to get these API credentials (note: the APIs associated web pages do change periodically, so the information below may not be up-to-date).

   - **Reddit:** Does not need API
   - **Twitter**. To access the Twitter API, you need to have a Twitter developer account. When logged into Twitter, then go to https://developer.twitter.com/en/portal/petition/use-case to create a Twitter developer account. Select “Academic” and “Student”. You will have to supply a description on how the app will be used. Try to provide as much detail as possible, as Twitter developer accounts are increasingly difficult to obtain, and there can be a lengthy approval process. After agreeing to terms and conditions and (if you haven't already done so) and verifying your email, your application will be reviewed. Go to “Project & Apps” and create a new project. Within the project, create a new app.

     Go to the page for your app, and on the apps’ page, click on the “Keys and tokens” tab. Generate and save a copy of the “Consumer keys” (API key and API secret key) and the “Access token & secret” keys. They will need to be supplied to vosonSML. For more on Twitter apps, see the Twitter Developers Site.

- **YouTube.** To access the YouTube API, you need to **use a non-Northwestern google account** (e.g. @gmail.com). When logged into Google, then go to the Google APIs Console and **create a project (regardless of whether you already have one)**. Then on to the navigation menu (top LHS) go to the APIs & Services page. Click on "+ ENABLE APIS AND SERVICES" → search for YouTube Data API v3, and click enable after you have pulled it up. This API should then appear in the Enabled APIs tab. Then, on the home menu, click on the. API & Services page. Click + Create Credentials and generate a Public API access key (choose "web server"). The API key then needs to be supplied to vosonSML.

3. Follow the instructions below to run the lab code in the lab, depending on whether you want to use the code we provided, or whether you would like to write your own code:

| **Those who work on the lab using the provided R code:** | **Those who work on the lab without the provided R code:** |
|---|---|
| Open the "Lab1_SocialMedia.R" file (this is a R script) in your R. If you use RStudio, from the menu, select "Session" → "Set Working Directory…" → "To Source File Location." This allows you to set your current working directory. | Create a new R script. Set a working directory. Load the following R pakcages 'magrittr,' 'igraph,' and 'vosonSML.' If your R doesn't have these packages, install these to R first. Then, load them. You can check if these packages are loaded in R by running 'sessionInfo()'. |
| Run the code we provided for the lab. We recommend that you run the code one line at a time, paying attention to what each line of code is doing and observing the output in the R console. | Create an "actor" network 'igraph' object from the data that you collected. The actor graph means that nodes are users and edges are based on replies/retweets/@mentions on Twitter and commenting on YouTube and Reddit. Also, store it as an igraph object meaning that your network graph should be recognized as 'igraph' when you run 'class(yournameofgraphhere)'. |

4. Check how many nodes and edges exist in the network.
Make sure that your network includes **at least 100 nodes**. DO NOT collect data including more than 1,000 nodes, as it can slow down the lab's code substantially. To increase or decrease the number of nodes, use the command numTweets/maxComments or add new hashtags/urls to your data collection. You may need to look at multiple related Twitter hashtags / reddit Threads / Youtube videos to gather more data.

5. Save your R environment as 'Lab1_SocialMedia.RData'. Keep a copy of this data, to submit on

Canvas, as well as to use again for Lab 1b. To load the RData you saved into R, run load('Lab1_SocialMedia.RData'). Make sure that your working directory is appropriately set when running this command.

**Part I Data collection questions to answer in your assignment:**

1. **(5 points)** Provide a high-level overview of the hashtags/urls you included in the data collection. Why did you choose this collection of hashtags/urls? Was there a specific, overarching question - intellectual or extracurricular curiosity - that motivated this collection of hashtags/urls?
2. **(2 point)** What are the insights you hope to glean by looking at the network of hashtags/urls - in terms of individual node metrics, sub-grouping of nodes, overall global network properties?
3. **(2 point)** Is the graph directed or undirected?
4. **(2 point)** How many nodes and links does your network have?
5. **(2 point)** What is the number of possible links in your network?
6. **(2 points)** What is the density of your network?
7. **(5 points)** Briefly describe how your choice of dataset may influence your findings. What differences would you expect if you use different hashtags/urls?

# PART II: Network Visualization (15 points)

In this part, using the data you are collecting, you will visualize the network and interpret these visualizations. Include a copy of the network plots you generate in your assignment.

**Instructions for Part II:**

Complete the following my modifying the code we provided, or by writing your own code.

1. Calculate the number of components in your graph.
2. If you have more than one component in your graph, create a giant component graph from your graph (pull out the data from only the single largest component in the network to visualize). Calculate the number of nodes and edges in the giant component graph.
3. Plot the giant component graph. Try different options to make it nicer than the default plot. Change node size, node color, and edge arrow size at least. Refer to this manual for more info on plot options.
4. Plot the giant component graph using a different graph layout option for the second visualization, adjusting other plot options as needed to obtain a decent visualization.

**Part II questions + plots to answer in your assignment:**

1. **(5 points)** Create a visualization of the whole network and include it in your report (the first visualization). In a paragraph, describe the macro-level structure of your graph based on the visualization. Is it a giant, connected component, are there distinct sub-components, or are there isolated components? Can you recognize common features of the subcomponents? Does this visualization give you any insight into the interaction patterns of your topic? If yes, what? If not, why?
2. **(5 points)** Create a second visualization, now using only the single largest component of the network (i.e., "giantGraph" if you work with the provided R code) and include it in your report. Are there any differences between the first visualization and second one? If so, why? If not, why not?

   (*If your whole network already had only one component to start with, the first and the second plots should be very similar. This is ok. Explain why the visualizations are similar or slightly different.)*
3. **(5 points)** Create a third visualization using a different 'igraph' layout option and include it in your report. Experiment with visualization options to make your layout better or add additional information to the plot. Explain your choice of layout options. In a few sentences, describe what types of observations are easier to make using one plot or the other.