**MSIA\_490-0\_SEC25– FALL 2019**

**Descriptive Analytic Exercise 1: Visualizing and Interpreting Networks**

**Deadline:** **Tuesday, October 22nd, noon**

The purpose of this lab is to learn how to conduct descriptive network analysis using the statistical software package R. Using the “[vosonSML](https://cran.r-project.org/web/packages/vosonSML/index.html)” package, this assignment will make use of a data set you collect from one of three social media platforms (i.e., Twitter, YouTube, and Reddit) by defining either hashtags (e.g., #MeTooMovement) or urls (e.g., [a metoo movement thread](https://www.reddit.com/r/AskReddit/comments/7sk2ox/what_are_your_views_on_the_metoo_movement/) on Reddit and 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.

**General Instructions:**

1. This lab has four parts indicated with Roman numerals (I, II, III, IV) in the outline below.
2. Prepare a report that includes your responses to all the **questions** for your reports 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 **Tuesday, October 22nd, noon.**
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. We also recommend using R Studio: <https://www.rstudio.com/>
8. R Studio is not required, but it is a popular software application to use R. 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.

Data collection questions to address in your report:

1. **(5 points) Provide a high-level overview of the hashtags/urls you included in the data collection**
2. **(4 points) 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?**
3. **(1 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?**
4. **(1 point) Is the graph directed or undirected?**
5. **(1 point) How many nodes and links does your network have?**
6. **(1 point) What is the number of possible links in your network?**
7. **(2 points) What is the density of your network?**
8. **(5 points) Briefly describe how your choice of dataset may influence your findings.** **What differences would you expect if you use different hashtags/urls?**

Data Collection Instructions:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Those who work on the lab **without** the provided R code |  | Those who work on the lab **with** the provided R code |
| 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](https://www.vox.com/2018/4/5/17157240/me-too-movement-sexual-harassment-aziz-ansari-accusation)). 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. | 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](https://www.vox.com/2018/4/5/17157240/me-too-movement-sexual-harassment-aziz-ansari-accusation)). 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 | 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()’. | 2 | 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. |
| 3 | *If you choose to collect data from Twitter and YouTube, request an API key.* The following part of the instructions comes from Rob Ackland & Tim Graham, which is retrieved from <http://vosonlab.net/SocialMediaLab/access_API>)   * 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 account. When logged into Twitter, then go to the [Twitter Apps](https://apps.twitter.com/) site, and press the “Create an app” button. You need to fill in some information including the name of the app e.g. “BigDataCourse,” description, website, etc. You do not need to enter the “Callback URL.” 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) supplying a valid phone number in your Twitter profile, your app will be created. Access your app and go to the “Keys and Tokens” tab. The “Consumer API keys” (API key and API secret key) and the “Access token” & “Access token secret” will need to be supplied to vosonSML. For more on Twitter apps, see the [Twitter Developers Site](https://dev.twitter.com/). * **YouTube.** To access the YouTube API, you need to have a Google account. When logged into Google, then go to the [Google APIs Console](https://code.google.com/apis/console) and **create a project (regardless of whether you already have one)**. Then go to the navigation menu (top LHS) APIs & Services → click on “+ ENABLE APIS AND SERVICES” → search and enable YouTube Data API v3. This API should then appear in the Enabled APIs tab. Then go to APIs and Services → Credentials and generate a Public API access key (choose “web server”). The API key then needs to be supplied to vosonSML. | 3 | *If you choose to collect data from Twitter and YouTube, request an API key.* The following part of the instructions comes from Rob Ackland & Tim Graham, which is retrieved from <http://vosonlab.net/SocialMediaLab/access_API>)   * 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 account. When logged into Twitter, then go to the [Twitter Apps](https://apps.twitter.com/) site, and press the “Create an app” button. You need to fill in some information including the name of the app e.g. “BigDataCourse,” description, website, etc. You do not need to enter the “Callback URL.” 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) supplying a valid phone number in your Twitter profile, your app will be created. Access your app and go to the “Keys and Tokens” tab. The “Consumer API keys” (API key and API secret key) and the “Access token” & “Access token secret” will need to be supplied to vosonSML. For more on Twitter apps, see the [Twitter Developers Site](https://dev.twitter.com/). * **YouTube.** To access the YouTube API, you need to have a Google account. When logged into Google, then go to the [Google APIs Console](https://code.google.com/apis/console) and **create a project (regardless of whether you already have one)**. Then go to the navigation menu (top LHS) APIs & Services → click on “+ ENABLE APIS AND SERVICES” → search and enable YouTube Data API v3. This API should then appear in the Enabled APIs tab. Then go to APIs and Services → Credentials and generate a Public API access key (choose “web server”). The API key then needs to be supplied to vosonSML. |
| 4 | Once you decide your topic and social media platform, write a script to collect data using vosonSML. Check out the [vosonSML github page](https://github.com/vosonlab/vosonSML) and [reference manual](https://cran.r-project.org/web/packages/vosonSML/vosonSML.pdf) for how to collect data. | 4 | Once you decide your topic and social media platform, run one of data collection scripts (i.e., Twitter, YouTube or Reddit) by replacing with your choice of hashtags/urls and API key (Reddit doesn’t require API key). |
| 5 | 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)’. | 5 |  |
| 6 | Check whether your network is directed or undirected, and how many nodes and edges exist in the network. | 6 | Check whether your network is directed or undirected, and how many nodes and edges exist in the network. |
| 7 | **Make sure that your network includes at least 200 nodes. DO NOT need to collect data including more than 1,000 nodes unless you want. To increase or decrease the number of nodes, use the command numTweets/maxComments or add new hashtags/urls to your data collection.** | 7 | **Make sure that your network includes at least 200 nodes. DO NOT need to collect data including more than 1,000 nodes unless you want. To increase or decrease the number of nodes, use the command numTweets/maxComments or add new hashtags/urls to your data collection.** |
| 8 | Calculate network density of your network | 8 | Calculate network density of your network |
| 9 | Save your R environment as ‘Lab1\_SocialMedia.RData’ so that you don’t need to collect data next time. To use this RData you saved, run load(‘Lab1\_SocialMedia.RData’). Make sure that your working directory is appropriately set when you run the command. | 9 | Save your R environment as ‘Lab1\_SocialMedia.RData’ so that you don’t need to collect data next time. To use this RData you saved, run load(‘Lab1\_SocialMedia.RData’). Make sure that your working directory is appropriately set when you run the command. |

**PART II: Network Visualization (20 points)**

In this part, using your network based on the data collection, you will visualize the network and interpret it.

Network visualization questions to address in your report:

1. **(5 points) Plot a visualization of your network (i.e., “actorGraph” if you work with the provided R code) and include the image in your report. In the visualization, what do you find interesting and what are salient characteristics?**
2. **(5 points) 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?**
3. **(5 points) Create a second visualization using your giant component graph (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?**
4. **(5 points) Create a third visualization using another ‘igraph’ layout option and include it in your report. Explain your choice of layout option. In a paragraph, what similarities and differences do you observe between the second visualization and third one?**

Visualization Instructions:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Without** the provided R code |  | **With** the provided R code |
| 1 | Plot your 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](https://kateto.net/network-visualization) for plot options. | 1 | Plot your graph. Try different options to make it nicer. Refer to [this manual](https://kateto.net/network-visualization) for plot options. |
| 2 | Calculate the number of components in your graph. | 2 | Calculate the number of components in your graph. |
| 3 | If you have more than one component in your graph, create a giant component graph from your graph. | 3 |  |
| 4 | 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](https://kateto.net/network-visualization) for plot options. | 4 | Plot the giant component graph. |
| 5 | Calculate the number of nodes and edges in the giant component graph. | 5 |  |
| 6 | Plot the giant component graph using a different graph layout option from the second visualization. | 6 | Plot the giant component graph using a different graph layout option from the second visualization. |

**PART III: Individual Network Properties (20 points)**

In this part, you will compute individual-level network measures and identify some key users in your network. Further, you will conclude this lab exercise with discussion in your main findings based on the visualizations, measures and your own analysis.

Individual network properties questions to address in your report:

**(8 points) Include the diagram for your chosen centrality measure from part 4 of the** Individual Network Properties Instructions **below, and also provide a table ranking the top 5-10 nodes for each centrality measure. Each centrality means (a) in-degree, (b) out-degree, (c) betweenness, (d) in-closeness, (e) out-closeness, (f) eigenvector, (g) Burt’s network constraint, (h) hub score, and (i) authority score.**

1. **(8 points) Briefly describe each centrality measure. How is each computed and what does its number mean in your network (e.g., a high centrality score means…)?**
2. **(4 points) How does the centrality of nodes vary with different types of centrality metrics? Why is this the case? Please offer some potential explanations using certain nodes as examples.**

Individual Network Properties Instructions

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Without** the provided R code |  | **With** the provided R code |
| 1 | Create a ‘network’ graph object from the graph (i.e., an ‘igraph object’ of your giant component graph). To do this, you need to load ‘sna’ and ‘network’ packages. Once you created the ‘network’ object, you should run ‘detach(“package:igraph”) to unload ‘igraph’. Also, to check whether you successfully created the ‘network’ graph object, run ‘class(yournameofnetworkobjecthere)’. | 1 |  |
| 2 | Calculate the instructed centralities of each actor in the graph, except for network constraint, hub and authority (see below). | 2 | Calculate the instructed centralities. |
| 3 | To calculate network constraint, hub and authority, use ‘igraph::’ to call igraph commands. And, using the igraph object of the giant component graph, compute network constraint, hub and authority. | 3 |  |
| 4 | Plot the ‘network’ graph object using ‘gplot.target’ command in ‘sna’ package. The command produces a target diagram, centering by one of your choice of centralities. Your plot looks like [the plot of betweenness](http://melissaclarkson.com/resources/R_guides/documents/gplot_layout_Ver1.pdf) on the page 10 | 4 | Plot the target diagram of the graph. |

**PART IV: Global Network Properties (40 points)**

In this part, you will identify global network structures of your network such as subgroups within a network provides much information to social network researchers, and a variety of algorithms have been developed to identify and measure subgroups. You will use some of igraph’s built-in tools to identify subgroups and central nodes for visual inspection.

Global network properties questions to address in your report:

1. **(3 points) Briefly describe (a) what k-core is, (b) what insight this k-core decomposition method provides, and (c) how many k’s exist in your network.**
2. **(3 points) Visualize your network using k-core decomposition and include the visualization in your report. In a paragraph, discuss your interpretation of the visualization and whether the results of k-core decomposition make sense based on your expectations of the network.**
3. **(3 points) Pick one of community detection algorithms. Which community detection algorithm did you choose and why?**
4. **(3 points) How many communities have been created? For your network, what might a community of nodes potentially have in common?**
5. **(3 points) What is a modularity score? Interpret the modularity score of your results of community detection?**
6. **(6 points) Plot the clusters and include the plot image in your report. What information does this layout convey? Are the clusters well-separated, or is there a great deal of overlap? What differences are there between actors in the same cluster and across clusters? Describe the brokers between any components and cliques. What are common features of these brokers?**
7. **(3 points) Present and interpret the in- and out-degree distribution based on your network as well as a log-log plot. Compute and interpret the estimate of the c slope.**
8. **(3 points) Present in a plot the observed and simulated values for each average path length and clustering coefficient based on the original network and 1,000 randomly shuffled networks**
9. **(3 points) Based on these data would you conclude that the observed network demonstrates small world properties? If so, why? If not, why not?**
10. **(10 points) In two or three paragraphs, discuss your major findings of your network based on all the analyses you’ve done in this exercise and also your own additional analysis if necessary. Your answer here will be evaluated based on depth and comprehensiveness. Thus, you’re encouraged to utilize extra information to answer this question. For instance, you can take a look at your original data (i.e., “twitterData,” “youtubeData,” or “redditData” if you work with the provided R code) in R. These data frames include additional user, text, and time information for your network. Similarly, if you need more insights from your network, feel free to run correlation and regression analysis based on your data collection.**

Global Network Properties Instructions

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Without** the provided R code |  | **With** the provided R code |
| 1 | Calculate the coreness of each node in the giant component graph (You will use the giant component graph from now on. It is call “the graph” in the following.). | 1 | Run k-core decomposition of the giant component graph (You will use the giant component graph from now on. It is called “the graph” in the following.). |
| 2 | Plot the graph and color nodes based on the number of coreness. | 2 | Plot the graph and color nodes based on the number of coreness. |
| 3 | Run a community detection algorithm for the graph. Check how many communities are created. | 3 | Run a community detection algorithm for the graph. If you want to use another algorithm, replace with your choice. Check how many communities are created. |
| 4 | Calculate the modularity score of the community detection. | 4 | Calculate the modularity score of the community detection. |
| 5 | Plot the graph using the community detection results. Make sure that your nodes are colored based on the communities. | 5 | Plot the graph using the community detection results. Make sure that your nodes are colored based on the communities. |
| 6 | Plot the giant component graph using a different graph layout option from the second visualization. | 6 | Plot the giant component graph using a different graph layout option from the second visualization. |
| 7 | Create a plot for in-degree distribution of the graph. | 7 | Examine the in-degree distribution after running the provided command. |
| 8 | Create a log-log plot based on the in-degree distribution. Note that since log doesn’t take 0, add 1 to every in-degree. | 8 | Examine the log-log plot of the in-degree distribution after running the provided command. |
| 9 | Calculate a power law fit to the in-degree distribution. Hint: use ‘power.law.fit()’ in ‘igraph’ | 9 | Read the comments in the provide R code to figure out the output of power.law.fit(). |
| 10 | Create a plot for out-degree distribution of the graph. | 10 | Examine the out-degree distribution after running the provided command. |
| 11 | Create a log-log plot based on the out-degree distribution. Note that since log doesn’t take 0, add 1 to every in-degree. | 11 | Examine the log-log plot of the out-degree distribution after running the provided command. |
| 12 | Calculate a power law fit to the out-degree distribution. | 12 | Examine the output of power.law.fit() |
| 13 | Compute the clustering coefficient and the average path length for the graph. Also, compute the clustering coefficient and average path length for 1,000 randomly reshuffled networks based on the graph. Plot the distribution of 1,000 simulated clustering coefficient values from the reshuffled networks and add the vertical line on the plot indicating the value of average path length from the graph. Create the same plot for average path length | 13 | Compute the clustering coefficient and the average path length for the graph. Also, compute the clustering coefficient and average path length for 1,000 randomly reshuffled networks based on the graph. Plot the distribution of 1,000 simulated clustering coefficient values from the reshuffled networks and add the vertical line on the plot indicating the value of average path length from the graph. Create the same plot for average path length |
| 14 | Run a one-tail t-test to examine whether the value of clustering coefficient from the graph is different from the simulated distribution. Run the same t-test for average path length. | 14 | Run a one-tail t-test to examine whether the value of clustering coefficient from the graph is different from the simulated distribution. Run the same t-test for average path length. |
| 15 | For the final question, look at the original data collection file before you made an ‘igraph’ object. The data file contain more detail information than you’ve so far been using in this lab. | 15 | For the final question, look at “twitterData,” “youtubeData,” or “redditData” depending on your choice of the social media platforms. These data contain more detail information than you’ve so far been using in this lab. |