Carlos Sathler | cssathler@gmail.com
INFO-I 535 – MGMT ACCESS USE BIG DATA (Fall 2016)

PROJECT B REPORT EXTRA CREDIT OPTION 1

Lists all sources of help that you consulted

In addition to the resources provided by the class (project instructions and tutorials) I also used the suggestion from colleague Tanmoy Datta Choudhury who posted a link to a Youtube video in Piazza (see url in the last paragraph in this section), with a tutorial on how to filter nodes in Gephi - https://www.youtube.com/watch?v=UrrWA_t1rjc

For the analysis of the book I chose for the "Minimal Plus" part of the project - The Sorrows of Young Werther, by Goethe - I consulted web page with list of the book characters:
http://www.gradesaver.com/the-sorrows-of-young-werther/study-guide/character-list

For the analysis of the book I chose for the "Extra Credit Opion 1" part of the project – Anna Karenina - I consulted web page with list of the book characters:
http://www.sparknotes.com/lit/anna/characters.html

The idea of looking for sites with lists of characters for the books also comes from colleagues in Piazza url https://iudatascience.soic.scholargrid.org/courses/course-v1:iudatascience+I535-I435-B669+FALL_2016/a523f04dd8664a04b622c74c49dba476/.

MINIMAL (Part 1 of 3)

1. **Is a window of size 15 a good window size for the characters that you think are related?**

I think the window is too high. It created too many connections between characters and therefore too many edges between nodes that ended up having higher degree than necessary. In the end, because Gephi offers an option to filter out nodes with low degree, it was a big issue.

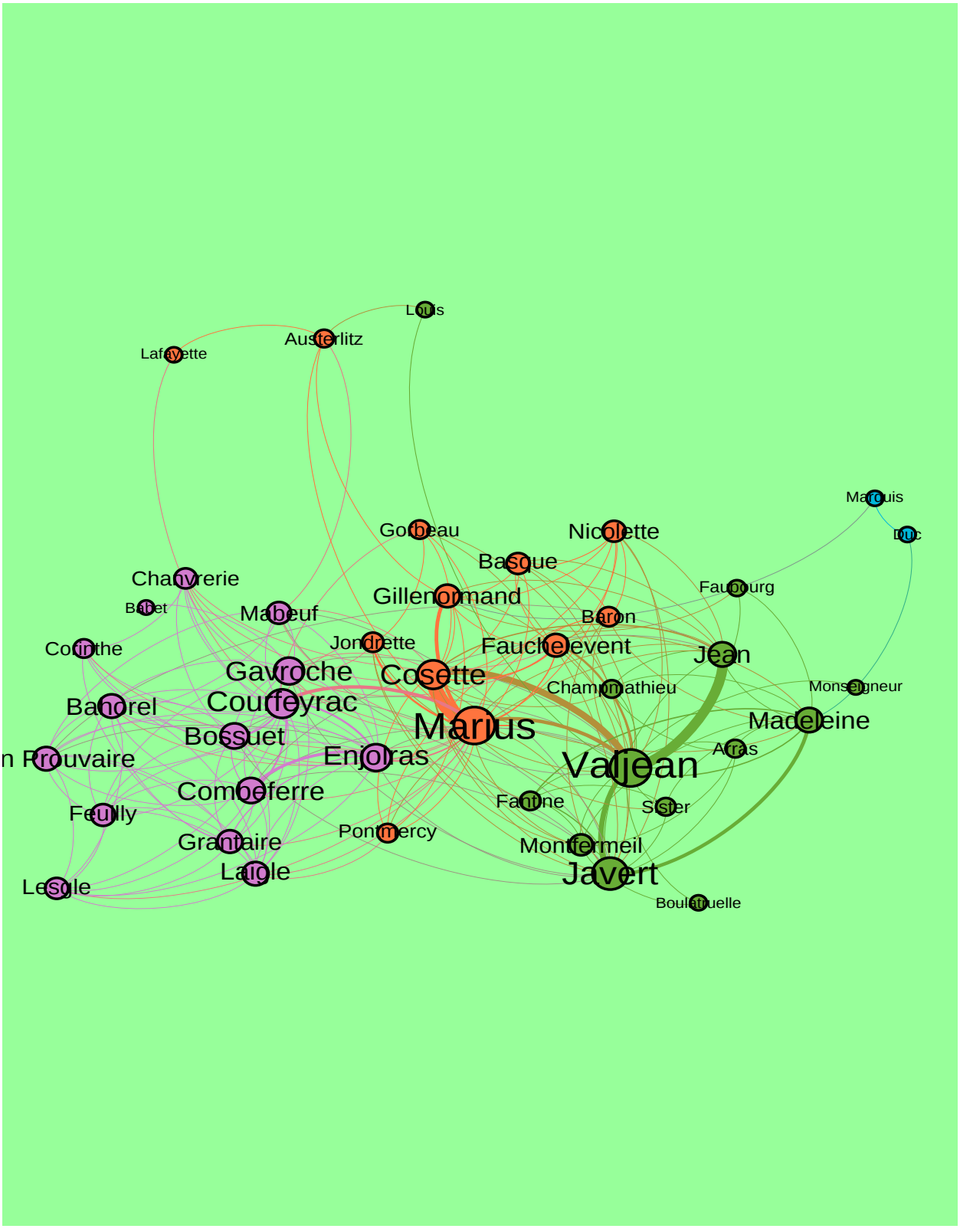
2. **What are the strengths and weaknesses of a larger window size? Give an example of a relationship that was missed because of a window size of N=15**

By using a small window size we could miss relationships between characters; conversely, a large window size may create a graph with too many edges and associate characters that are not meaningfully connected by the plot of the book. It may also augment the relatedness between characters. The combination of too many edges and overall higher degree for all nodes may end up polluting the graph and users may miss important relationships among the characters.

Given the total number of characters in any text, say C, and given N, the window size, any time N < C, all relationships missed are examples of a missed relationship for window = N. The only way to guarantee no relationships will be missed in any text is to make N >= C. In the graph that I produced, seen on next page, an example of a missed relationship, among many, is Javert and Pontmercy at the middle bottom of the graph.

3. Include a copy of the network graph (or portion of it) that you generated for the characters in Les Miserables from Gephi (PDF)

See graph below and Canvas (les-mis.pdf).



Carlos Sathler | cssathler@gmail.com
INFO-I 535 – MGMT ACCESS USE BIG DATA (Fall 2016)

MINIMAL PLUS (Part 2 of 3)

1. **When you analyzed texts of your own choosing that you're familiar with or interested in, did you glean any insights from this type of analysis that would be harder to glean from a simple read-through?**

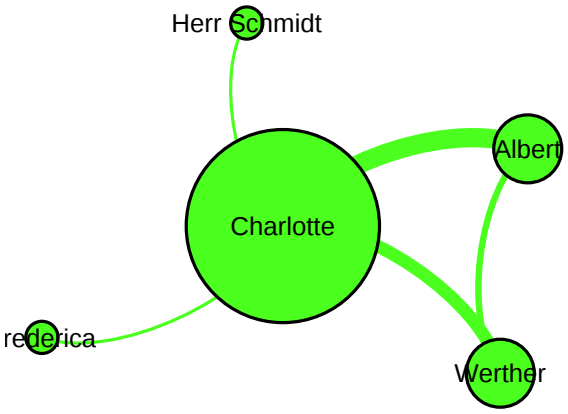
Absolutely. The Sorrows of Young Werther (by Goethe) is a classic celebrated story of unrequited love between the main character, Werther, and Charlotte for whom he falls madly in love, but who is promised to another man – Albert. The story speaks of a deep passion and the reader's attention is drawn heavily to the relationship between Werther and Charlotte. It was very surprising to see that the edge linking Charlotte to Albert signals a stronger relatedness between these two characters than the one between Charlotte and Werther. Goethe manages to get the reader absorbed in the deep connection between Werther and Charlotte, but clearly at the same time he crafts a concrete even stronger association between Charlotte and Albert. This apparent contradiction is in line with the sad (for Werther) ending of the story.

2. **Include a copy of the graph (or portion of it) that you generated for the characters in content you chose (PDF)**

See graph below and Canvas (werther.pdf).

3. **An archive containing the text(s) you chose to analyze (ZIP).**

werther.zip uploaded to Canvas.



Carlos Sathler | cssathler@gmail.com
INFO-I 535 – MGMT ACCESS USE BIG DATA (Fall 2016)

EXTRA CREDIT OPTION 1 (Part 3 of 3)

1. **When you extract the characters, create the network representation and apply the network analysis algorithms, there is some fine-tuning of the algorithms that needs to happen. Try exhaustively cleaning your list of characters, adjusting the parameter values for the length of the text window, or the number of communities.**

   How do the results differ? Merging nodes will reduce number of nodes with low degree and it will eliminate pairs of nodes with thick edges, as authors tend to use variations of the name of the character if they refer to the same character close in the text. Fewer nodes with higher degrees will also draw attention to main characters if node size attribute is mapped to node degree.

   Did you need to do a lot of fine-tuning to produce a visualization that was useful and easy to understand? Yes, mostly having to do with merging nodes after initial cleanup, and playing with the layout of the graph for clarity. It was also important to find a good low bound degree number for the degree filter as well as the proper resolution when running the modularity statistic routine, in order to yield an appropriate number of modularity classes (we settled with 5 for the final visualization). The detailed steps we followed on the project are listed below.

   What ways of automating this fine-tuning can you think of?

   A python program could be written to load characters of the book from a website or other source and parse the Chars variable that we created looking for ways to merge characters before creation of gml file for load into Gephi. It would be helpful if Gephi would have a plug that users could use to create Macros, like MS Excel macros.

   a) These were the high level/summary steps I performed to clean data:

      Ordered nodes in alphabetical order and manually deleted the following nodes: God, French, every node with "&#" in the label, Young, Whom, Which, Woman, Venus, Thou, Thy, Thy Truth, Thy Whole Covenant, Tit, Thank God, Stay, Sorry, Socialism, Please, Plaisir, Old, Often, Naples, Milan, Merciful, Meine, Louis Quinze, Leo Tolstoy, Lent, Leave, Jesus Christ, Instinct, Him, Hence, Has, Half, Gladiator, French, Frenchman, Frenchwoman, Fleurs, Flung, Europe, Doubt, Dolly, Does, Divorce, Dinner, Cord, Christianity, Christ, Baden, Burgundy, Brownie, Bravo, Between, Better, Beside, Babylon, Atlas, Arseny, Amuse, Almighty God.

      Ran Statistics "Average Degree" and looked for extra nodes with high degree that could be deleted. Deleted historical figures: Alexander Nevski

      Obtained list of characters in the book from website http://www.sparknotes.com/lit/anna/characters.html. Used the list to merge characters in Gelphi: Agafya Mikhailovna, Alexander Kirillovich Vronsky, Alexei Alexandrovich Karenin, Alexei Kirillovich Vronsky, Anna Arkadyevna Karenina, Countess Vronsky, Darya Alexandrovna Oblonskaya

(Dolly), Ekaterina Alexandrovna Shcherbatskaya (Kitty), Elizaveta Fyodorovna Tverskaya (Betsy), Fyodor Vassilyevich Katavasov, Konstantin Dmitrich Levin, Landau, Madame Stahl, Marya Nikolaevna, Nikolai Dmitrich Levin, Nikolai Ivanovich Sviyazhsky, Prince Alexander Dmitrievich Shcherbatsky, Princess Shcherbatskaya, Sergei Alexeich Karenin (Seryozha), Sergei Ivanovich Koznyshev, Stepan Arkadyich Oblonsky (Stiva), Varvara Andreevna (Varenka), Varvara Vronsky, Vasenka Veslovsky, Yashvin

b) Tuning Graph for Effective Visualization
    a. Added filters: Giant Component + Degree Range (chose 5 <= degree <= max degree)
    b. Enabled labels and ran layout routines ForceAtlas 2 (scaling = 50), NoOverlap and Label Adjust
    c. Ran statistics Average Degree, Network Diameter, Modularity (resolution = 1 yielded 7 groups, which is not bad)
    d. Sized nodes based on degree and colored groups
    e. Graph is polluted…
    f. Changed filter for degree range (15 <= degree <= max degree)
    g. Ran statistics modularity with higher resolution = 2 and got 5 groups
    h. Ran layout routines Contraction and repeated step b and d
    i. Went back to resolution = 1 and 7 groups for a richer graph
    j. Repeated step h, this time using ForceAtlas 2 with scaling = 150
    k. Noticed that "Sergei" and "Sergei Ivanovitch" had thick edge and must be same character; the same for the following pairs: "Alexey" and "Alexey Alexandrovich"; "Stepan Arkadyich" and "Arkadyich"
    l. In the Data Laboratory merged "Sergei" and "Sergei Ivanovitch"; "Stepan Arkadyich" and "Arkadyich"; "Alexey", "Alexey Alexandrovich" and "Alexey Alexandrovich Karenin"
    m. Changed degree filters to 12 <= degree <= max degree
    n. Reran all statistics and all layouts a few times
    o. Manually grouped nodes with same modularity class (color)
    p. Ran layout Adjust Label one last time
    q. Created PDF

2. Include a copy of the graph (or portion of it) that you generated for the characters in content you chose (PDF)

Please see on next page and also Canvas (karenina.pdf).

3. An archive containing the text(s) you chose to analyze (ZIP).

karenina.zip uploaded to Canvas.

Carlos Sathler | cssathler@gmail.com