CS 432/532 Web Science: Assignment #4

Alexander C. Nwala

Bethany DeMerchant

9 March 2019

Contents

Background:

The Friendship Paradox is a social networking phenomenon which says that an average individual's friends have more friends than the individual. This report demonstrates simple tests of this phenomenon performed using one account each on Facebook, Twitter, and LinkedIn:

> Determine if the friendship paradox holds for [a given account]. Compute the mean, standard deviation, and median of the number of friends that… friends have. Create a graph of the number of friends (y-axis) and the friends (x-axis) themselves, sorted by number of friends (y-axis).

Background (ii):

Examination of the data quickly revealed that most data sets had a few friends with a dramatically higher number of connections than the other friends in the data set. For the sake of academic interest, z-scores were computed for each friend in the data set and those data points more than two standard deviations from the mean were omitted from a subset X_2. The mean, median, and standard deviation were then calculated for X_2.

Methods— General

For each data set, the number of friends the user being examined has includes the user themselves. The number of friends any given friend of the user has does not include that friend but does include the user. Numbers are rounded to two decimal places where necessary.

For each set of original data, one graph is presented with the y-axis in linear scale and one is presented with the y-axis in log scale. Horizontal lines have been placed at $y = \bar{x}$, $y = \overline{x} \pm s$, and $y = \overline{x} + 2s$ on the linear scale graphs[1]. For each set of modified data, one graph is presented with the y-axis in linear scale and horizontal lines have been placed at $y = \bar{x}$, $y = \overline{x} \pm s$, and $y = \overline{x} + 2s$. All graphs are generated using the Matplotlib module for Python.

Methods— Facebook:

Facebook requires a user to approve collection of information regarding their friend list, preventing the collection of live data. Instead, analysis was performed on the 2014 friend data of user `Alexander Nwala`, provided as a file `acnwala-friendscount.csv`. This file is provided in the `A4/Data` folder, along with a `facebook.txt` file containing the same data.

Analysis— Facebook:

`Alexander Nwala` has 99 Facebook Friends. 87 of these Friends have more Friends than him. The mean for this sample is 538.19, the standard deviation is 538.52, and the median is 395. `Alexander Nwala`'s 99 Friends is well below the mean and median of the data set. As predicted, it is shown that most of `Alexander Nwala`'s friends have more friends than him as of the 2014 sample.

[1] As $\overline{x} - 2s < 0$ in all cases, the line that would mark $y = \overline{x} - 2s$ has been omitted to avoid cluttering the graphs.

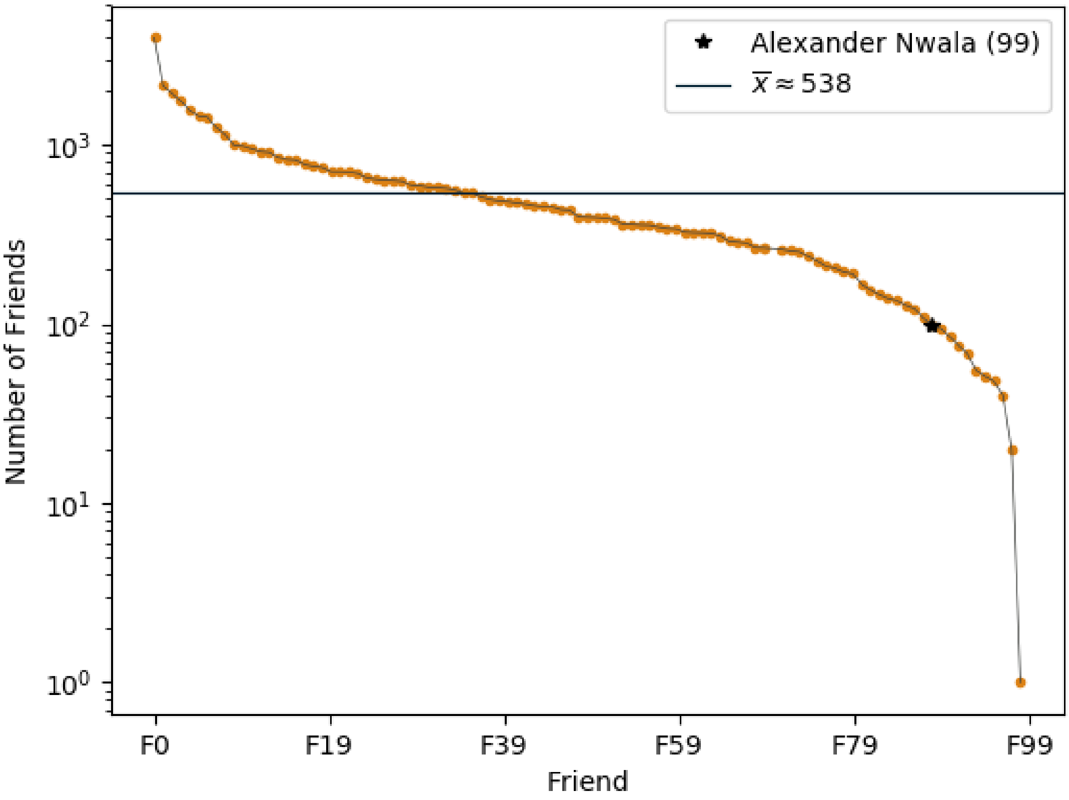

Figure 1: Facebook Friends, Log Scale

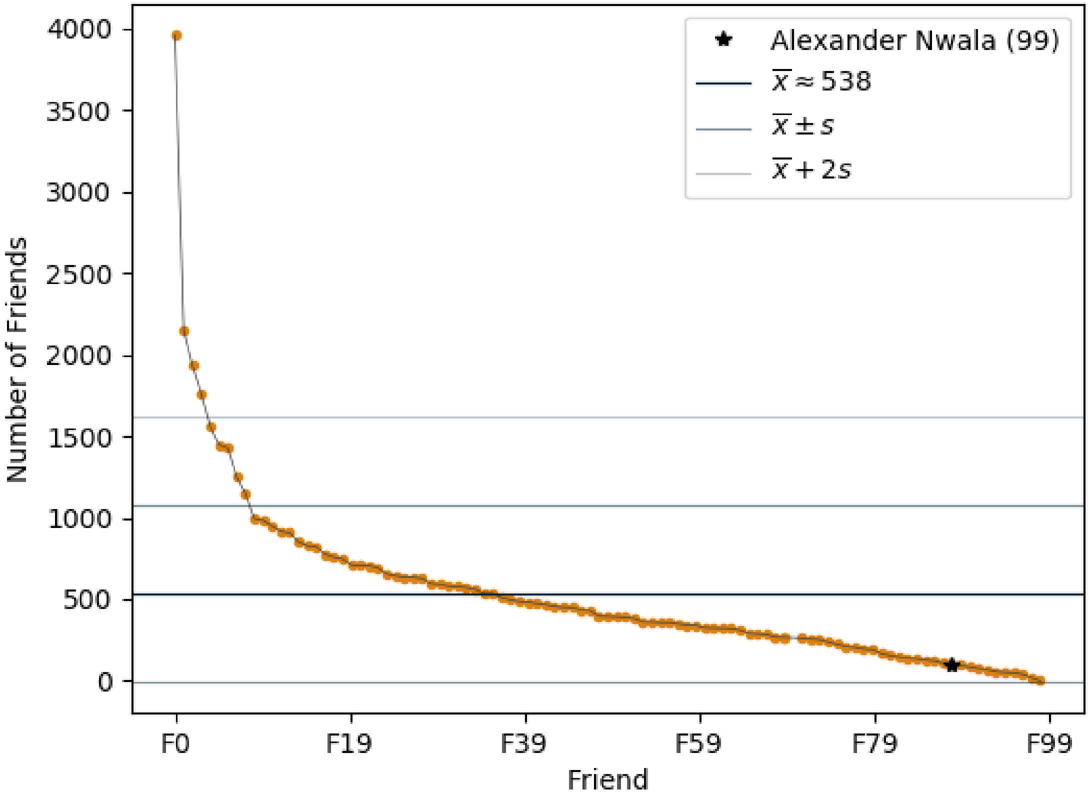

Figure 2: Facebook Friends, Linear Scale

Analysis— Facebook (ii):

As seen in Figure 2, four data points are more than two standard deviations from the mean: Friends with 3995, 2143, 1931, and 1757 Friends respectively. Excluding these data points for further analysis results in a sample of 95 Friends, with a mean of 457.84, standard deviation of 328.31, and median of 389. Of these 95 Friends, 83 have more friends than `Alexander Nwala`. Using these numbers, `Alexander Nwala` has fewer than the mean number of Friends and is a full standard deviation from the mean.

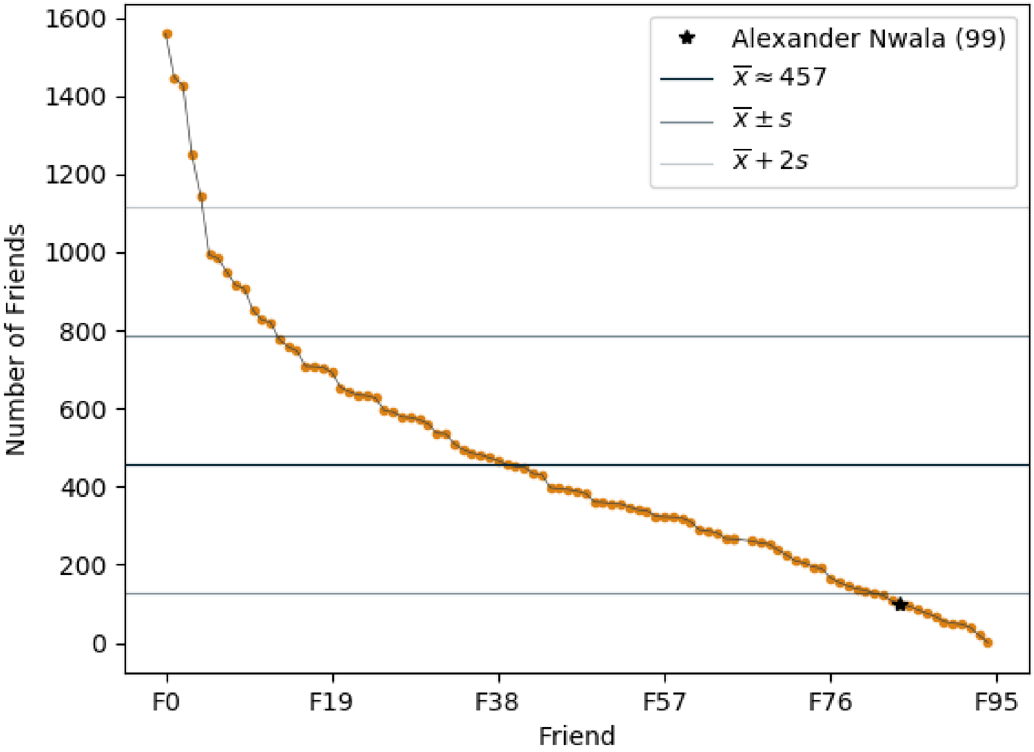

Figure 3: Facebook Friends, Normalized

Methods— Twitter (followers):

Data for Twitter account `acnwala` was collected using the Twitter API and `Tweepy` library for Python. The list of `acnwala`'s followers was downloaded using

```
flist= limit_handled(tweepy.Cursor(api.followers, screen_name='acnwala').items())
```

where `limit_handled()` is a method provided by Tweepy[2] to prevent problems with the Twitter rate limit. The `tweepy.Cursor` object is used both to support the `limit_handled()` method and to facilitate collection of all followers rather than only the first twenty. The following code was then used for each follower to create a list of all followers and the number of followers each follower has:

```
user = api.get_user(follower)
foff_list.append((user.screen_name, user.followers_count))
```

The data was then saved to `tw_follows.txt`, available in the `A4/Data` folder.

While not strictly necessary for processing the data, associating each follower's screen name with their follower count was helpful in organizing and proofing the collected data as well as the script itself.

[2] http://docs.tweepy.org/en/3.7.0/code_snippet.html#handling-the-rate-limit-using-cursors

Analysis— Twitter (followers):

The Twitter account `acnwala` has 249 followers. The mean for this sample is 1698.91, the standard deviation is 7414, and the median is 278. 131 of 249 followers have more followers than acnwala; most of them have several times more.

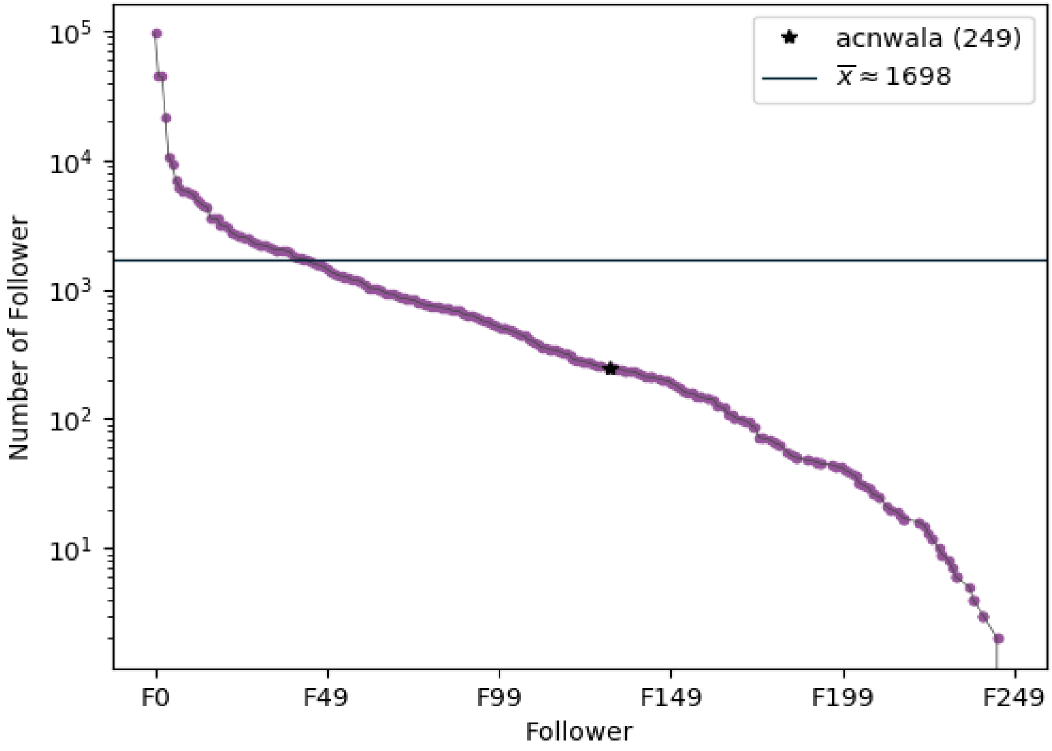

Figure 4: Twitter Followers, Log Scale

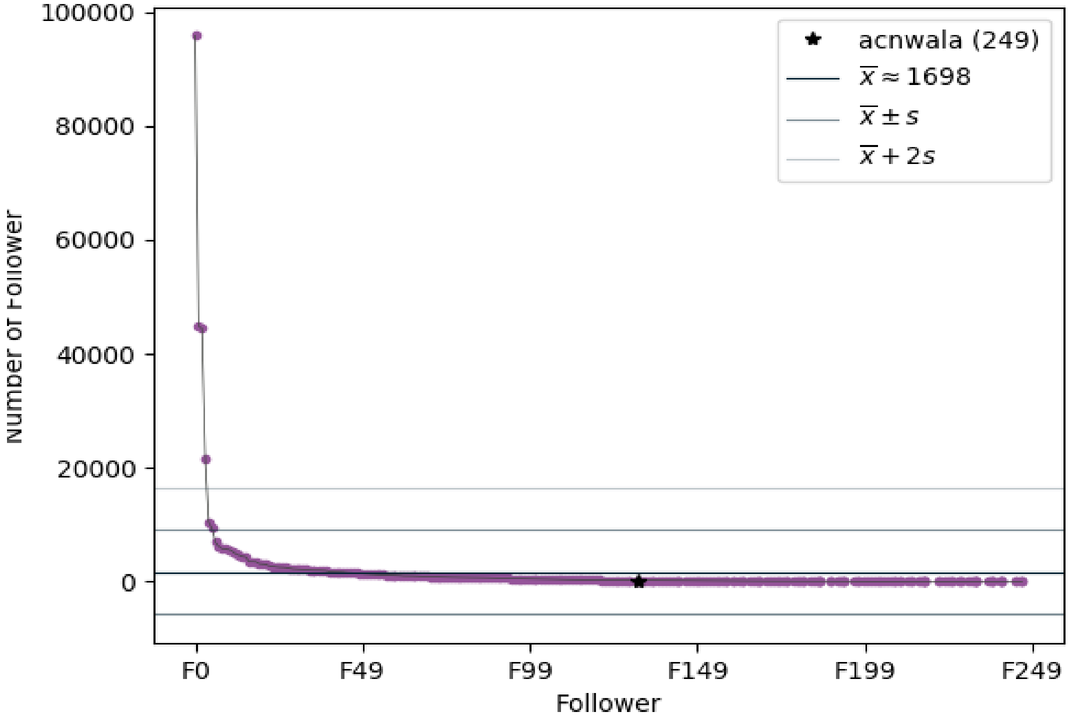

Figure 5: Twitter Followers, Linear Scale

Analysis— Twitter (followers) (ii):

The data on acnwala's followers is strongly affected by a few very high values. Four followers have follower counts of 95811, 44845, 44419, and 21539 respectively.[3] Removing the outliers from the analysis leaves a sample of 245 followers, 128 of whom have more followers than acnwala. The mean of this sample is 883.33, and the standard deviation is 1492.54. The median is 271.

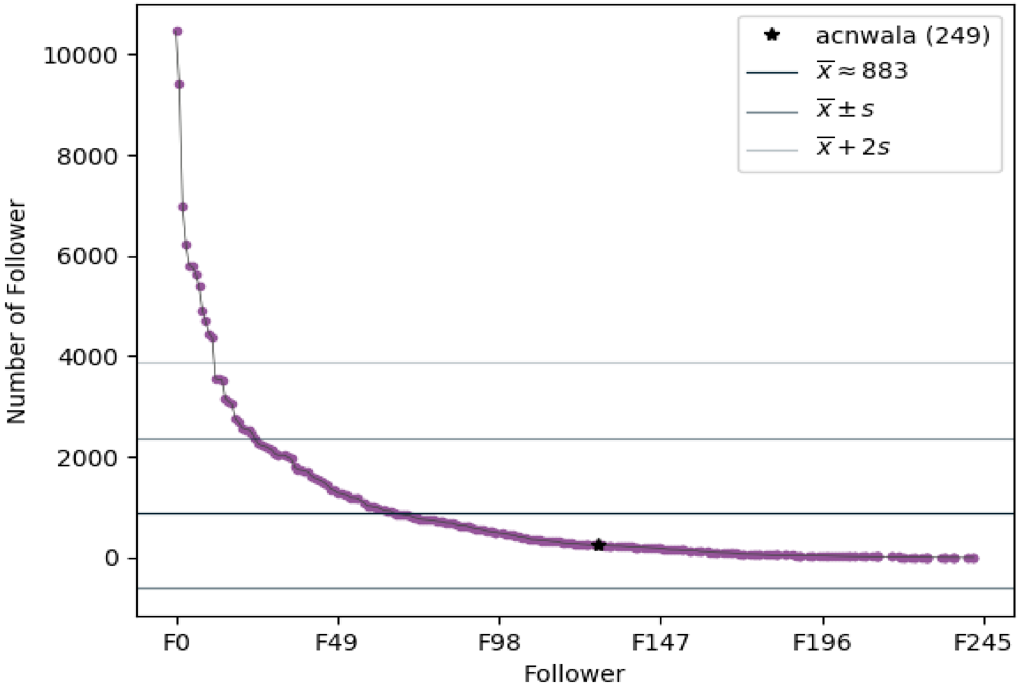

Figure 6: Twitter Followers, Normalized

Methods— LinkedIn:

Analysis was performed with permission on the connections data of user Traci DeMerchant. To avoid complications with the API such as experienced with Twitter, the data was collected by hand.[4] The collected data is stored in a file linkedin.txt, provided in the A4/Data folder.

Analysis— LinkedIn:

LinkedIn user Traci DeMerchant has 44 connections. The mean for this sample is 346.77, the standard deviation is 326.67, and the median is 199. 42 of 44 connections have more connections than Traci DeMerchant; most of them have several times more.

[3] For scale, the next highest value is 10,459.

[4] In other words, my mother was kind enough to log in to her LinkedIn profile and let me count her connections.

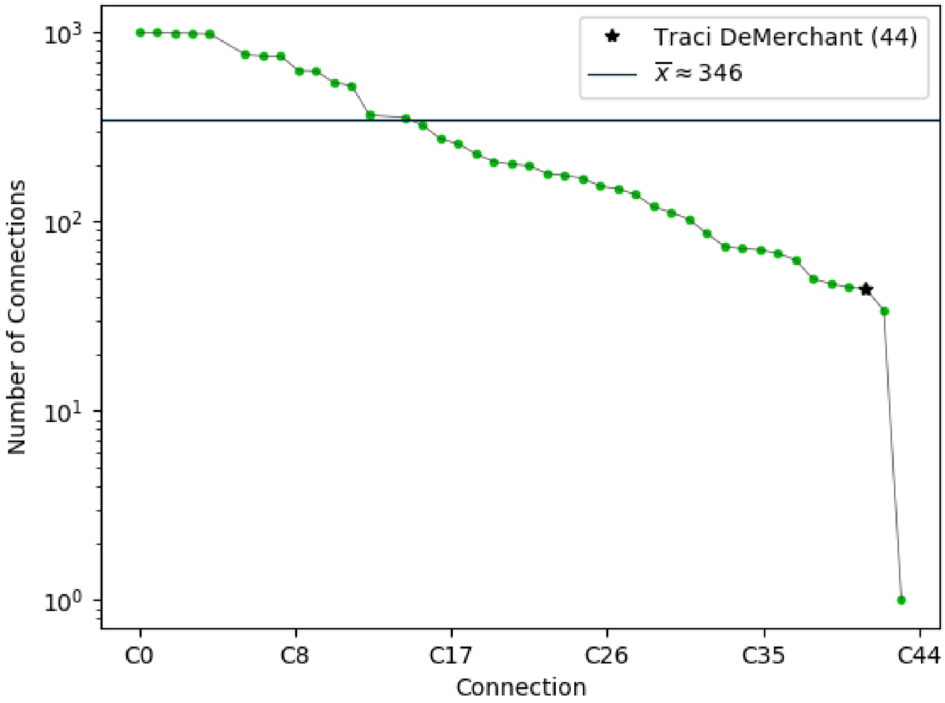

Figure 7: LinkedIn Connections, Log Scale

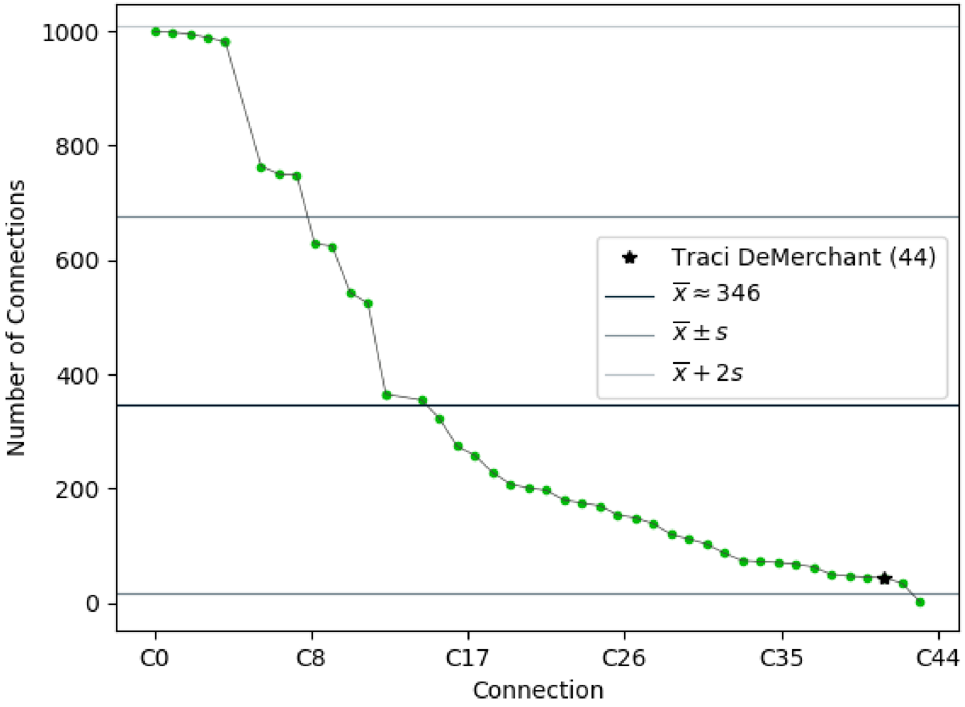

Figure 8: LinkedIn Connections, Linear Scale

Analysis— LinkedIn (ii):

No values in the sample of LinkedIn connections are more than two standard deviations from the mean. However, one value corresponds to an account that is not used and has no connections other than `Traci`

`DeMerchant`[5]. Discarding this value as an outlier has only a small effect on the data. The mean of the remaining 43 values is 354.81, and the standard deviation is 326.52. The median is 201.

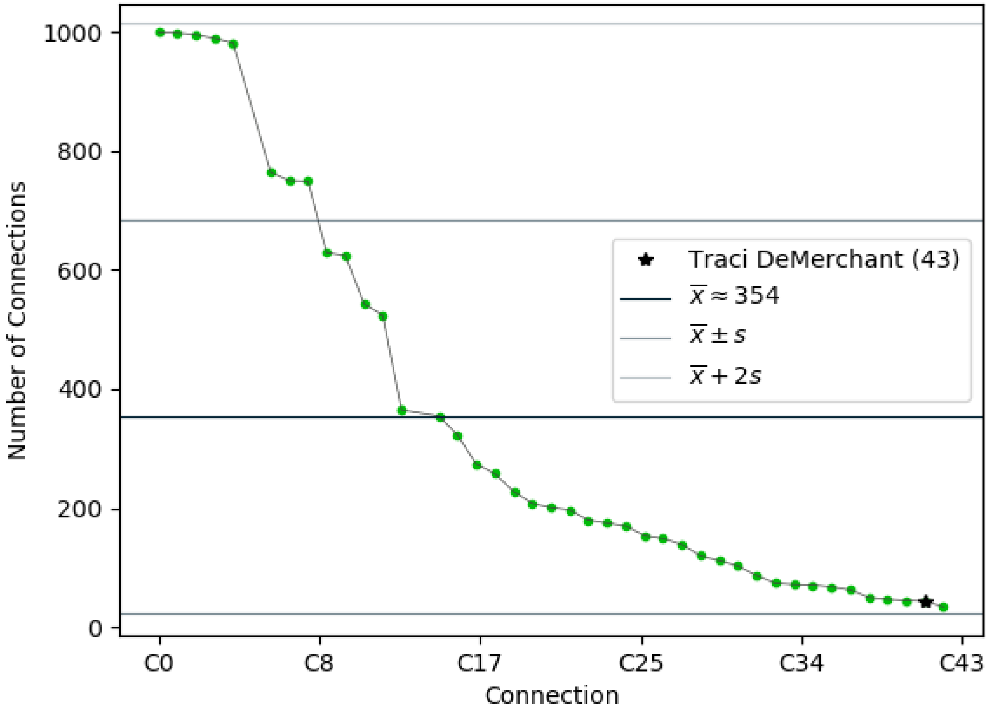

Figure 9: LinkedIn Connections, Normalized

Methods— Twitter (friends):

The Twitter API uses the term `friends` to refer to those users a Twitter user is following[6]. The same term is used in this report for clarity. Analysis was performed on friends data of `acnwala` using the same methods outlines under Methods— Twitter (followers), with the line

```
flist= limit_handled(tweepy.Cursor(api.friends, screen_name='acnwala').items())
```

used to collect `acnwala`'s friends data and the code

```
        user = api.get_user(friend)
        foff_list.append((user.screen_name, user.friends_count))
```

used to create a list of friends and their friend counts. The data was then saved to `tw_friends.txt`, available in the `A4/Data` folder.

Analysis— Twitter (friends):

The Twitter account `acnwala` has 93 friends. The mean for this sample is 1185.04, the standard deviation is 2594.07, and the median is 450. 79 of 93 followers have more followers than `acnwala`.

[5] I made a profile for the specific purpose of viewing data for this assignment, so it isn't a "real" LinkedIn profile.

[6] https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-friends-ids

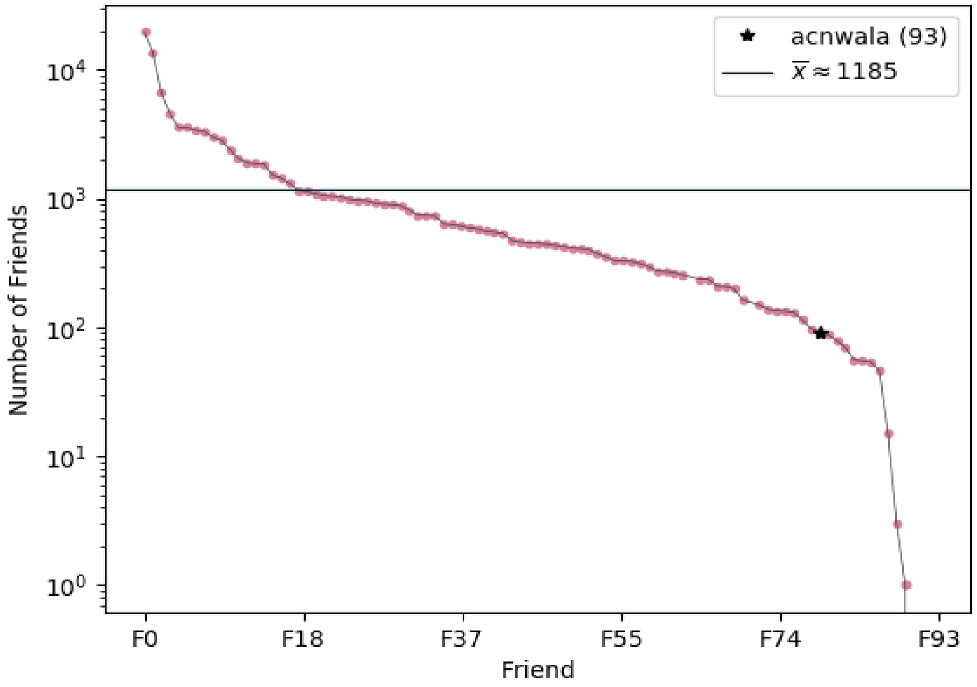

Figure 10: Twitter Friends, Log Scale

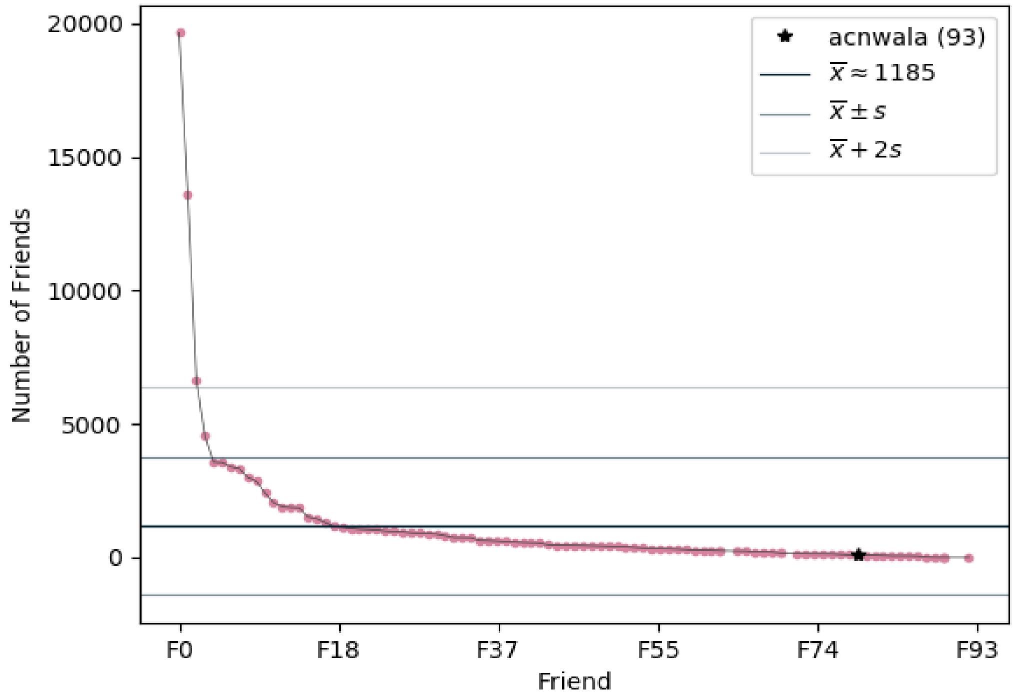

Figure 11: Twitter Friends, Linear Scale

Analysis— Twitter (friends) (ii):

The data on `acnwala`’s followers is strongly affected by a few high values. Removing these values from the analysis leaves a sample of 90 friends, 76 of whom have more friends than `acnwala`. The mean of this sample is 781.17, and the standard deviation is 944.74. The median is 440.

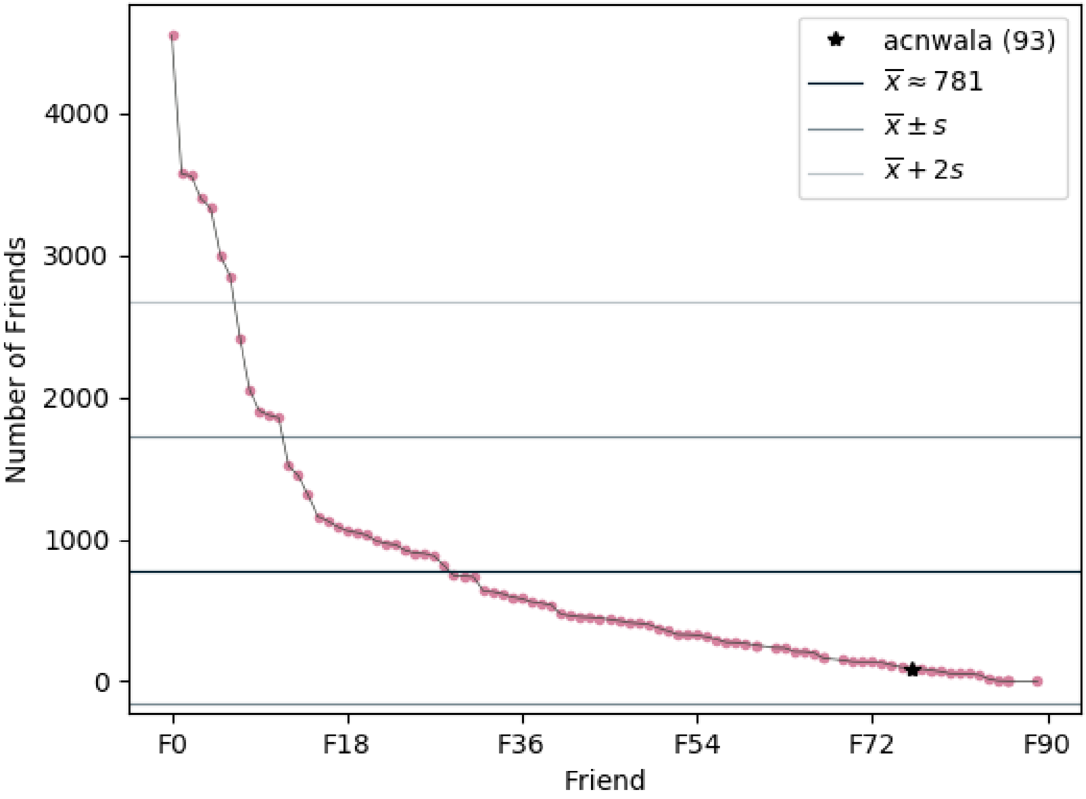

Figure 12: Twitter Friends, Normalized