Comparing the Size of Friend Groups Between Western and Central Europe using Twitter API Data

Pierre Winter
ETH Zurich, Switzerland

Abstract—**Does the size of friend groups change geographically across different countries? This project uses social network data to validate Dunbar's number and to investigate differences in the size of friend groups between Western and Central Europe. Public Twitter timeline data is accessed from the Twitter API and the definition of what constitutes a friend is also defined. The author uses a two-sample t-test as well as permutation tests as the test statistics and finds some statistically significant differences between the number of Twitter friends in the two geographical regions. Additionally, the size of the friend groups are found to be much smaller than Dunbar's number of 150 friends.**

I. Background

A. Dunbar's Number

Studies have shown that there is a correlation between an individual's brain structure and the size of their social group [1]. More specifically, the size of the neocortex in mammals is linearly correlated with the size of their social groups. Larger social groups require a larger neocortex because they present cognitive constraints as well as time costs that are required to maintain these relationships [2]. Dunbar has proposed that the maximum number of friends a person can actively maintain in their social group is approximately 150, which the social data science community has deemed Dunbar's number. This finding has also shown to be consistent even with the advent of online social networks [3], which some have hypothesized could allow for humans to hold more than 150 friends. In this work, social network data will be used to further validate Dunbar's number and to test for noticeable differences between the number of friends for two human populations.

B. Twitter API

Data from the Twitter social network was used in this work for both theoretical and practical reasons. It is a fundamentally useful tool to study social interactions because of its widespread use in all layers of society. Among the popular social networks, it is also currently one of the more generous with respect to applying for and receiving an access token to interact with their API. This project was performed using the R software environment with the *rtweet* package. This package facilitates both accessing the Twitter API and parsing the returned data directly into an R data frame.

II. Methods and Data Handling

A. Data Structure

In order to address the project questions, Twitter user timeline data was required across Western and Central Europe. The countries constituting these regions are defined in this work as:

Western Europe	**Central Europe**
Portugal	Denmark
Spain	Netherlands
France	Germany
Italy	Austria

Twitter timeline data is straightforward to access with the *get_timeline()* function in the *rtweet* package. With additional access to the Google Maps API, the *get_timeline()* function also allows for timeline retrieval based on city coordinates or names. The capital city names for each country in the table above were used in this work as these cities are assumed to have the largest population and population density of Twitter users. An example of an API call for 200 Twitter timelines using *rtweet* for the capital city of Portugal, Lisbon, is given by:

```
search_tweets(n=200,
    geocode=lookup_coords(address='lisbon'))
```

The data returned by the API is the corresponding timeline information presented in an R data frame for the 200 most recent tweets. This process of retrieving Twitter timelines was repeated for all 8 capital cities such that a total of 1600 timelines were requested. While *rtweet* does much of the data cleaning, it can also return tweets or timelines which are not publicly available. This private data is removed from our analysis and we then subsample the remaining public data such that 200 unique Twitter timelines from Western Europe and 200 unique Twitter timelines from Central Europe remain.

B. *Defining Friendship*

For each of the 400 unique Twitter users in our core sample, a way of measuring friendship needs to be defined. Here we define a user's friend as someone who has reciprocated (both to and from) a Twitter mention greater than or equal to n_{thresh} times in their h most recent Tweets. n_{thresh} and h are tunable parameters that we can set based on empirical testing and they will have a direct impact on the perceived number of Twitter friends. For example if $n_{thresh} = 4$ and $h = 200$, *User A* and *User B* are considered friends only if *User A* mentions the user ID of *User B* 4 or more times and *User B* mentions the user ID of *User A* 4 or more times within each of their last 200 tweets.

In this project $n_{thresh} = 2$ and $h = 100$. This was found to be a good compromise between how friends may actually interact on Twitter and how much time is required for requesting a large number of Twitter timelines using *rtweet*. Upon finding the number of Twitter friends $n_{friends}$ for the 200 Western Europe and 200 Central Europe samples, it was found that the distribution of the number of friends was highly skewed as seen in Figure 1.

A large percentage of people had zero Twitter friends by our definition. Because many of the available statistical tests rely on data being approximately normally distributed, a log transform was applied to the number of friends:

$$n^*_{friends} = \ln(n_{friends} + 1) \tag{1}$$

where 1 was added such that the variable transformation is not undefined for users who had $n_{friends} = 0$. The inclusion of 1 extra friend is not expected to change the conclusions drawn from our analyses. This transformation moderately improves the normality of the variable distribution as seen in Figure 2. This transformed variable $n^*_{friends}$ will therefore be used throughout the rest of this work.

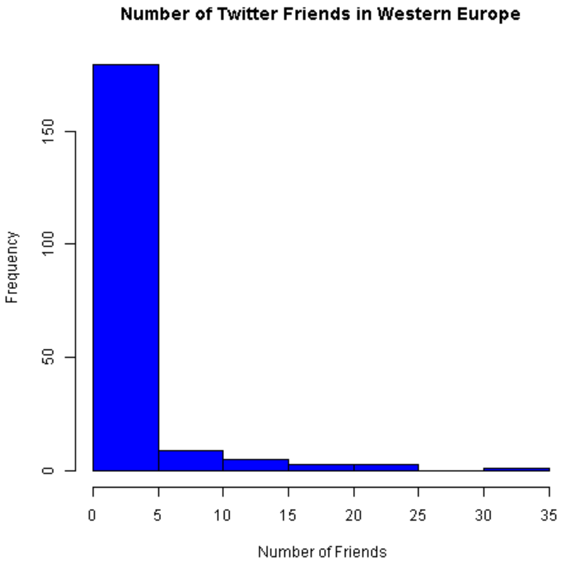

Fig. 1.

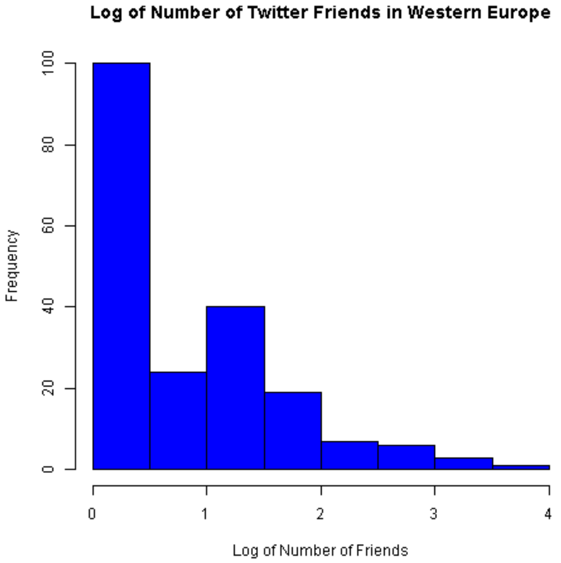

Fig. 2.

III. Statistical Analysis

A. Null Hypothesis

For a statistical test to be valid we need to define a null hypothesis with which to compare against. In the scope of this work, the null hypothesis is that there is no difference between the distributions of Twitter friends in Western and Central Europe. The alternative hypothesis which we will try to prove, thereby rejecting the null hypothesis, is that there is a difference between the distributions of Twitter friends in Western and Central Europe. The possible outcomes of this analysis could therefore be one of the following: 1) There is no difference in Twitter friends between the two regions 2) Western Europe has more Twitter friends than Central Europe 3) Central Europe has more Twitter friends than Western Europe 4) There is inconclusive evidence to accept or reject the null hypothesis. Three different statistical tests will be performed in order to arrive at one of these conclusions.

B. t-Test

An independent two-sample t-test, defined using the difference in means, will be analyzed first. The t statistic is defined here for the Western (W) and Central (C) European sample populations using Welch's individual variance [4]:

$$t = \frac{\bar{X}_W - \bar{X}_C}{\sqrt{\frac{s_W^2}{n_W^2} + \frac{s_C^2}{n_C^2}}} \tag{2}$$

$\bar{X}$ is the population's mean, s is the population's standard deviation, and n is the population's sample size. As shown in Figure 3, the t-value (red) was found to be -3.36 and this is compared against a t-distribution (black) centered at zero with a total of 398 degrees of freedom (400 samples - 2 means).

The p-value for this test is calculated to be 0.0008 which corresponds to the area under the t-distribution curve for values smaller than $t = -3.36$. A p-value of 0.0008 suggests that Central Europe has a significantly larger number of Twitter friends than Western Europe. It is important to note however that formal t-tests rely on the samples coming from normal distributions, which is not entirely the case here. A statistical test which is more robust for non-normal data will be investigated next.

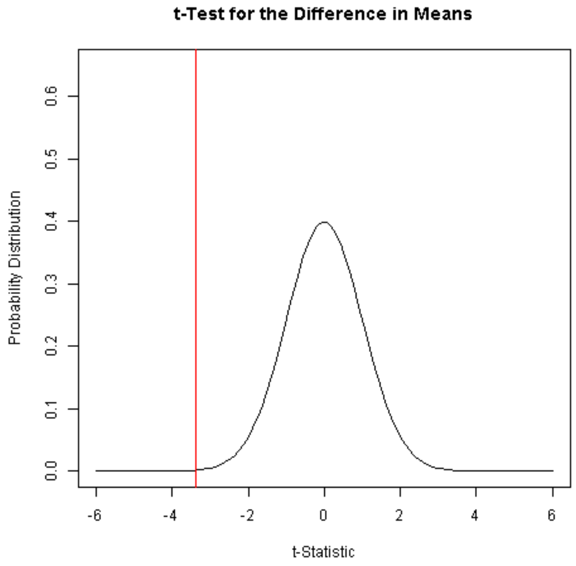

Fig. 3.

C. Permutation Test

A permutation test is more appropriate for the analysis here because it creates a baseline distribution without having to require that the samples are drawn from a normal distribution. To perform the permutation test, all 400 samples were placed together and their number of Twitter friends $n^*_{friends}$ were randomly shuffled to create new Western Europe and Central Europe distributions. A new statistic, the absolute difference in means $\bar{X}_W - \bar{X}_C$, was calculated for these newly shuffled distributions. This process was repeated 10,000 times and all corresponding values were aggregated to create a baseline distribution of random samples. This is shown in Figure 4 along with the absolute difference in means of 0.326 from the original samples Using the absolute difference in means as the test statistic results in a p-value of 0.0005 which again suggests that the Central Europe test population has a significantly larger number of Twitter friends than the Western Europe test population.

Because the data is highly skewed, it can be useful to perform the same permutation test as described above but with the absolute difference in medians as the test statistic. This is shown

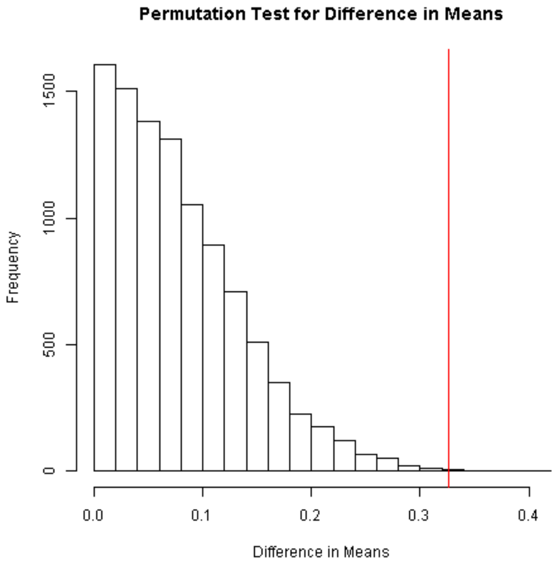

Fig. 4.

in Figure 5 along with the absolute difference in medians of 0.347 from the original samples Using the absolute difference in means as the test statistic results in a p-value of 0.16 which suggests that the null hypothesis cannot be rejected. The distribution of permuted data in Figure 5 is also much farther from normal than the difference in means data from Figure 4, which means that, the conclusions using this test statistic are perhaps not valid.

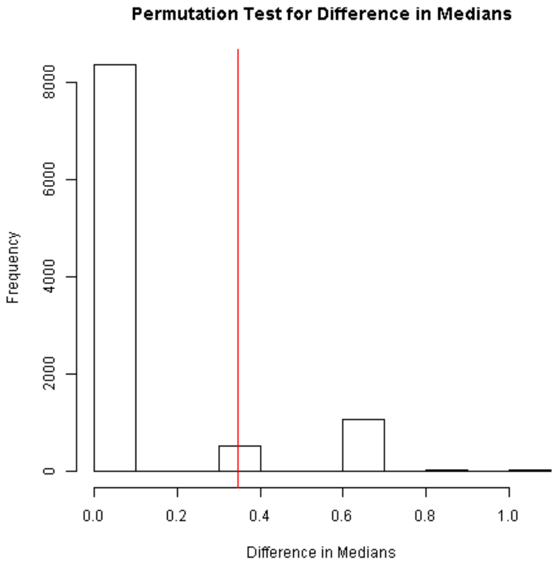

Fig. 5.

IV. Conclusion

Statistical tests were performed on public Twitter timeline data to determine the number of Twitter friends users have in Western and Central Europe. A two-sample t-test, as well as permutation tests for the absolute differences in means and medians, were then performed on the two European populations. It was found that the permutation test using the absolute difference in means was the most reliable test statistic and this concluded that users from Central Europe had more Twitter friends than users in Western Europe. Additionally, the size of the friend groups in both populations are found to be much smaller than Dunbar's number of 150 friends.

A. *Limitations*

This analysis has room for improvement, most notably with how a Twitter friend is defined. The threshold of reciprocated mentions n_{thresh} and the number of historical tweets to parse through h are the parameters which define a friendship in this study. These empirical parameters have a large effect on the resulting friend distribution and so should be tested more closely in future studies. The friendship definition used here does not analyze the content of the tweets nor the Twitter account. This means that some people who we consider Twitter friends could be engaged in arguments or they could even be public entities such as news agencies.

Another factor to consider is that users in different countries may use Twitter more or less than others based on historical or cultural factors. In addition, our sample of users may be biased because we only searched for user timeline from capital cities. Sampling from multiple cities for each country could give a better representation of Twitter users and their interactions.

B. *Future Outlook*

Next steps to improve this analysis would be to acquire more user timelines, which requires time and computational resources. With these resources, it would also be useful to increase the number of historical tweets to parse through h to more than 100. This would potentially increase the number of perceived Twitter friends for users who use

Twitter frequently. Furthermore, including more countries and cities within each country would also help make our samples more representative of the populations in Western and Central Europe.

REFERENCES

[1] J. Sallet, R. B. Mars, M. P. Noonan, J. L. Andersson, J. X. O'Reilly, S. Jbabdi, P. L. Croxson, M. Jenkinson, K. L. Miller, and M. F. S. Rushworth. Social network size affects neural circuits in macaques. *Science*, 334(6056):697–700, 2011.

[2] R. I. M. Dunbar. Do online social media cut through the constraints that limit the size of offline social networks? *Royal Society Open Science*, 3(1):150292, 2016.

[3] Bruno Gonalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on twitter networks: Validation of dunbar's number. *PLOS ONE*, 6(8):1–5, 08 2011.

[4] B. L. WELCH. THE GENERALIZATION OF STUDENT'S PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED. *Biometrika*, 34(1-2):28–35, 01 1947.