Putting abortion opinions into place: a spatial-temporal analysis of Twitter data

Amanda Jean Stevenson

Abstract

There is a great deal of uncertainty about how people feel about reproductive rights and how the distribution of that sentiment varies by location. Large scale surveys ask questions in ways that are highly susceptible to bias and they rarely give us specific geographical information on respondents' locations. Moments like Wendy Davis' filibuster of the Texas omnibus abortion restriction bill precipitate reactions and thus may bring latent, hard to measure sentiment to light. I use data from tweets about Davis' filibuster to describe the discourse on Twitter about the bill and Davis. Estimating users' locations using a variety of methods, I describe the spatial and temporal distribution of resistance to and support for abortion restrictions in Texas.

Introduction

While abortion is widely held to be a divisive and polarizing issue, the opinions of Americans on the issue are not well understood. There is substantial debate about whether or not polarization about abortion is increasing over time and while the red state versus blue state debate is prominent in our media discourse, the opinion dynamics within states are not described by this dichotomy. The present work examines the case of Twitter participation in a discussion of an omnibus abortion restriction bill around the time Wendy Davis' filibuster of the bill. The filibuster and bill received substantial state and national attention, with more than 180,000 people watching the live-feed of the Texas State Senate at one point during Davis' filibuster. Additionally, there was substantial political engagement surrounding the bill Davis was filibustering, with thousands of protestors returning to the Texas Capitol multiple times to lodge their opinions and represent their positions on the bill during hearings and debates.

This critical moment precipitated wide public discussion of abortion, presenting an opportunity to capture the distribution of latent opinions about abortion in a diverse but conservative state. Protestors used social media, and particularly Twitter, to communicate plans and rally support. Together with the wide engagement with the issue of abortion, the fact that Twitter was the primary organizing medium for individuals engaged in the protest provides social scientists with an opportunity to capture the distribution of opinion using individual-level data.

This paper relies on demography's longstanding focus on the social diffusion of demographic phenomena to describe how expressions of resistance to abortion restrictions spread. I use a very large social network dataset based on the 1.66 million tweets (from 277,315 individual users) about the filibuster to investigate the temporal-spatial diffusion of expressions of resistance to the bill and support for Davis' filibuster on the social network site Twitter. I also observe diffusion through social networks for a subsample of the users. Drawing on recent computer science methodologies for spatial location estimation (for example Davis et al 2011), I use a simple but innovative social network analysis to find locations for Twitter users who have not enabled GPS data. I also estimate users' locations using their declared location in their Twitter user profile.

State-level restrictions are the most prevalent form of new abortion restriction in the US and thus demographers with an interest in reproductive health can look to the role that social processes like diffusion play in building this resistance. Examining the processes leading to the diffusion of resistance or support for abortion restrictions may help demographers understand how demographically important policies like these have emerged and may emerge in the future.

This is particularly important as a growing number of states are instituting dramatic restrictions on access to reproductive health care, including abortion and reactions to these policy changes, particularly on social media, can influence governmental and private policy

decisions. For example, the outrage over the Komen Foundation's defunding of the Planned Parenthood breast health program led to a full replacement of the money.

More fully understanding participation in these expressions of resistance will include understanding the spatial distribution of their participants. The methods outlined here facilitate that understanding. It contributes substantively to our understanding of the spatial distribution of opinions about abortion and how the expression of those opinions evolved over the course of the debates over the omnibus abortion bill in Texas. It contributes methodologically by outlining and implementing a process by which social scientists can locate Twitter users in space, even when they do not enable GPS geolocation and by testing this method against other location estimation methods.

Methods

Work on this project has included data procurement, descriptive analysis of the content and temporal pattern of participation, the development of software for estimating Twitter users' locations, preliminary software implementation, and analysis of the findings from preliminary implementation.

Data Procurement

The objective of the present work is to describe the spatial variation in Twitter users who tweeted about Wendy Davis' filibuster and the omnibus abortion bill in Texas during Summer 2013. Therefore, I secured a comprehensive dataset of all tweets with hashtags involved with the debates surrounding the bill. I formulated the list of hashtags initially and then had the list validated by key informants with central roles in the resistance to the bill, including two bloggers and a professional social media coordinator.

Based on this list of hashtags, I purchased a comprehensive dataset of all tweets from the dates June 19th through July 14th, 2013 with any of the hashtags from the third party data warehouse TweetReach. The dates covered by the data include all major events involved in the

omnibus abortion bill's passage through the Texas legislature. June 19th was the day before the first large public hearing about the bill and July 13th was the day the bill finally passed. Wendy Davis' filibuster was June 25th.

Twitter users who were represented in the database of all tweets to relevant hashtags were considered target users whose location I developed an algorithm to find.

Descriptive Analysis

I analyze the temporal patter of use and initial use by hashtag in order to describe the scale and content of participation in the online resistance to the omnibus abortion bill.

Software Development for Spatial Estimation Based on Social Networks

The tweets themselves do not have locations associated with them. Therefore, I decided to use the Twitter API to collect location data and build location estimates based on social network estimation. The Twitter API is a feature of the Twitter website that allows direct access to some Twitter data from a computer.

The user data available from the Twitter API has two types of location data. It has cell phone latitude and longitude coordinates for some users and has a user generated location text string some users. Relatively few users have the latitude and longitude data (less than 5% in our sample analyzed thus far) and the text string location can be unreliable. For example, some users in our sample report their location as "between a rock and a hard place" and "Milky Way." However, additional location data are available using the tweet data, which include GPS location for each tweet from users who have enabled GPS. Unfortunately, relatively few users have enabled GPS. In our sample analyzed thus far, the proportion is about 7%.

Together with a software engineer, I designed software in Python and R to collect all available location data on all users who tweeted with the relevant hashtags and to use their social networks to estimate their locations when other location data were unavailable. The outline of this software is included in Appendix A. The method is informed by computer science research.

Roughly, the software establishes a database to hold the social network data for the users, calls the Twitter API to find GPS, cell phone data, and text string location information for the target user, all the target user's mutual friends, and all the mutual friends of the mutual friends of the target user. Then, using a voting algorithm, it assigns locations to the mutual friends of the mutual friends, and using the locations found for that group it assigns locations to the target user's mutual friends. With this set of estimated locations for the target user's mutual friends, the software then estimates the location of the target user. Because this is a new method for locating Twitter users, the calculation is performed for all target users without respect to their GPS, phone, or text string location and the results are compared to their GPS, phone, and text string locations. These comparisons are used to validate the method. See Table 1 for preliminary findings.

This software is computationally intensive and because of the size of the dataset a full analysis will require more than a billion calls to the Twitter API, which occupies the processor for several seconds for each call. Thus, the software has been tested for a small sample of the data on a desktop computer.

Once I have implemented the algorithm to secure reliable location data for the Twitter users who tweeted with the hashtags, I will examine maps of tweets by county and day and I will test for social contagion and spatial contagion (using Moran's I (Gatrell et al, 1996) and the methods outlined in Van Steeg's 2012 paper.

I will describe spatial tweeting as a density of tweets to the hashtags by county. This density will be the number of tweets to the collected hashtags divided by the number of tweets to the #txlege hashtag from the county.

Results

My spatial results are preliminary because my analysis requires time on a supercomputer and fully implementing the analysis will require permission to access greater volumes of data

from the Twitter API than are usually authorized. I am in the process of securing such permission from Twitter and my time on the TACC supercomputer array is scheduled for early 2014.

Analysis of the non-spatial component of the 1.66 million tweets reveals the temporal pattern of expressions of resistance to the abortion restrictions.

Figure 1. presents a panel of the number of tweets by hashtag by date. It illustrates that the most intense participation coincided with the dates of public hearings and debates. The dramatic spike is in the afternoon and evening of June 25th, when Davis filibustered the bill. The bill had four separate names during the period of analysis. The inclusion of other bill titles and their rough exchange of volume, with one becoming prominent as another recedes, illustrates the importance of including multiple hashtags when analyzing Twitter data, as the relevance of hashtags in identifying tweets changes over time.

It is important to note that the y-axis for the #standwithWendy and #HB2 hashtags is 10 times as large than the y-axis for the other hashtags. This is because at the time of the filibuster, and at the time of the bill's final passage, the volume of tweets per day for these hashtags was very high.

In order to illustrate the dynamics of entry into the Twitter discussion regarding the bill, Figure 2 displays the number of users first using any of the hashtags by date. This figure illustrates the dramatic increase in participants in the Twitter discourse with these hashtags on the day of Davis' filibuster.

Figure 3 displays the hashtags used by Twitter users at the time of their first tweet with any of the hashtags. These charts illustrate the changing points of entry into the discussion over time. At the time of Davis' filibuster, initial tweets were much more likely to include the #standwithWendy hashtag, whereas #standwithTXwomen became a point of entry around the time that a coalition of advocacy groups organized a large rally a few days after the filibuster.

Further work will display these line charts as stacked area charts to illustrate the changing composition of entrants' hashtag use.

In order to validate our spatial location algorithm, Table 1 displays the concordance between GPS-estimated locations, social location estimates, and text string location estimates for the sample that has run. While the pairwise agreement between the different methods of user location estimation is not very high (about 50%), the social location estimation process overestimates location in large cities. For users whose text location or frequent GPS location was a metropolitan area, the social location estimator performed much better – about 75% agreement in the preliminary data.

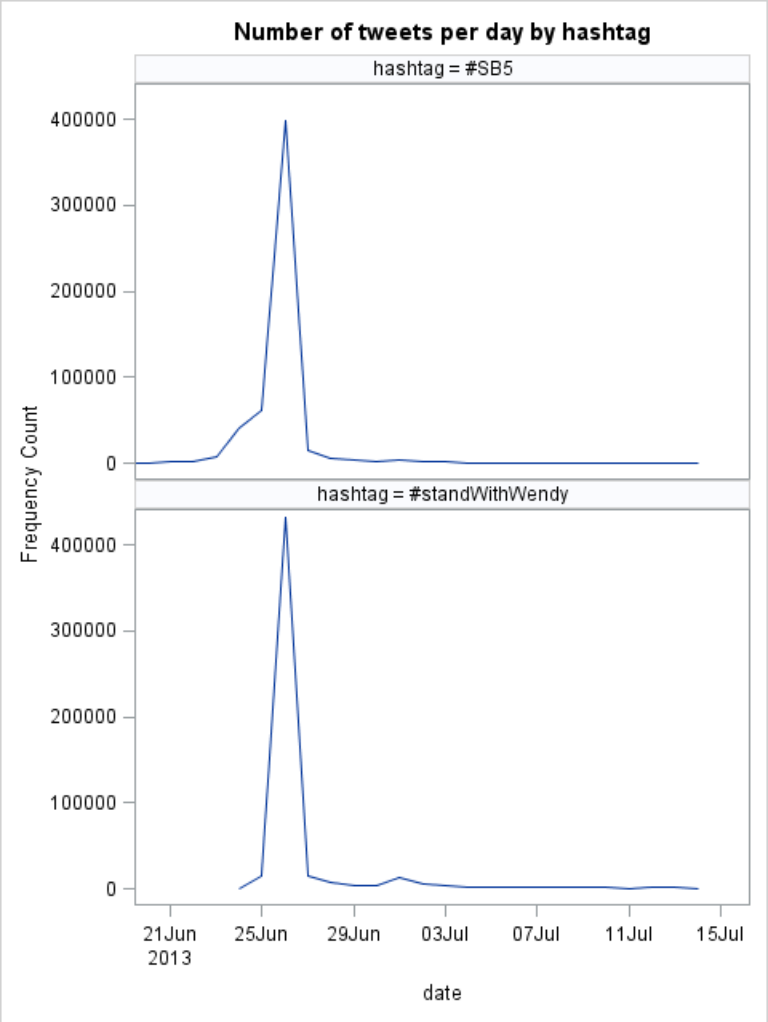
In the full paper, the most important results will be the time series of maps and the statistical tests for contagion. Unfortunately, I do not have sufficient location data to complete this component. However, the forgoing results demonstrate that the data are rich and informative even without this important element of the analysis.

Limitations

The most important limitation of this work is that Twitter use is not universal, and furthermore not all Twitter users discuss their opinions on abortion. Therefore, this analysis is not representative at the population level. However, Twitter use has become dramatically more prevalent and population representative in recent years (Pew 2011). What this work does do is describe the spatial distribution of participants in a popular discussion about abortion restrictions. The fact that individuals chose to become participants and used Twitter to do so presents less of a selection bias to the extent that Twitter the a primary form of communication and information dissemination during the debates and protests.

Figure 1. Total tweets with hashtags by date

Panel A: Total tweets with #standwithWendy and #SB5 by date



Panel B. Total tweets with bill name hashtags by date

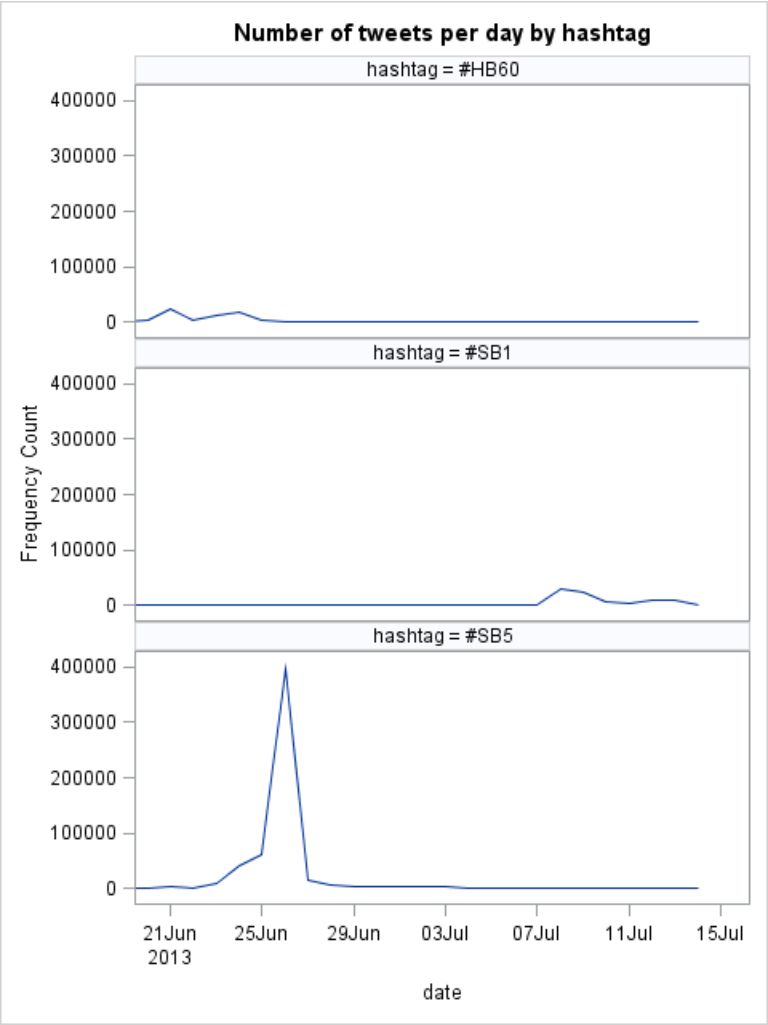


Figure 2. Number of Twitter users first tweeting with any of the hashtags by date

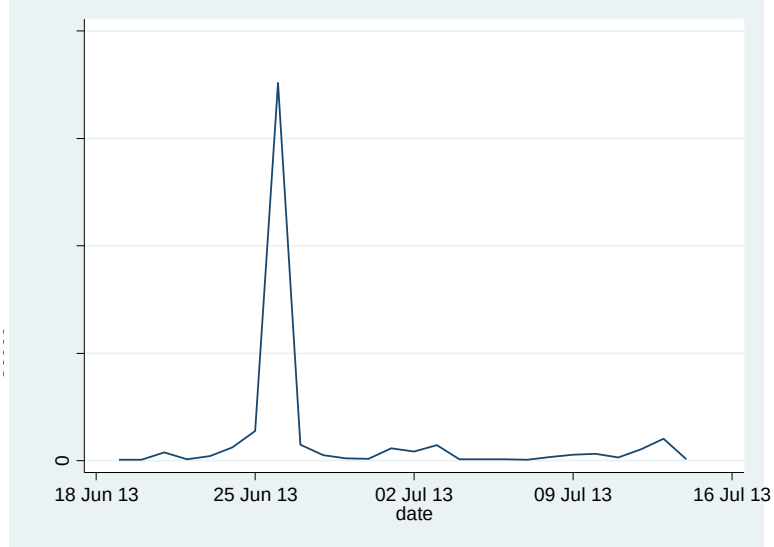
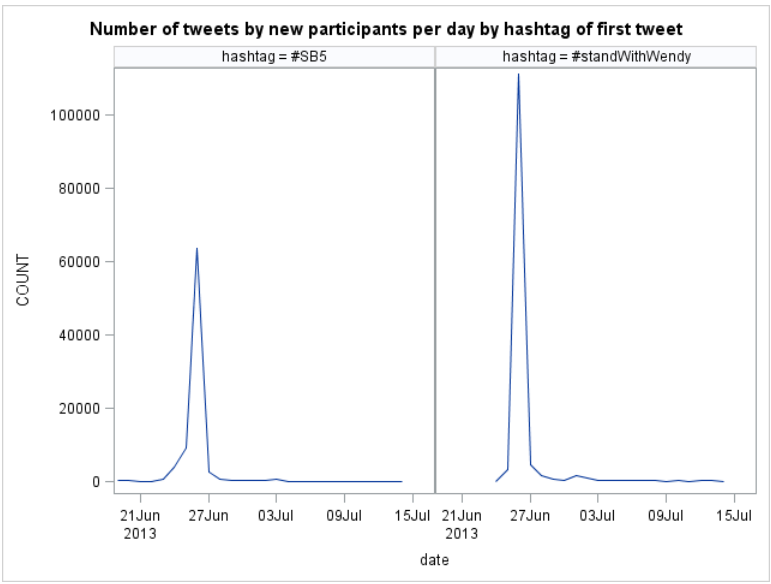


Figure 3. Number of users first using any of the hashtags by date

Panel A. Entry tweets using most common entry hashtags



Panel B. First time tweets using less common entry hashtags

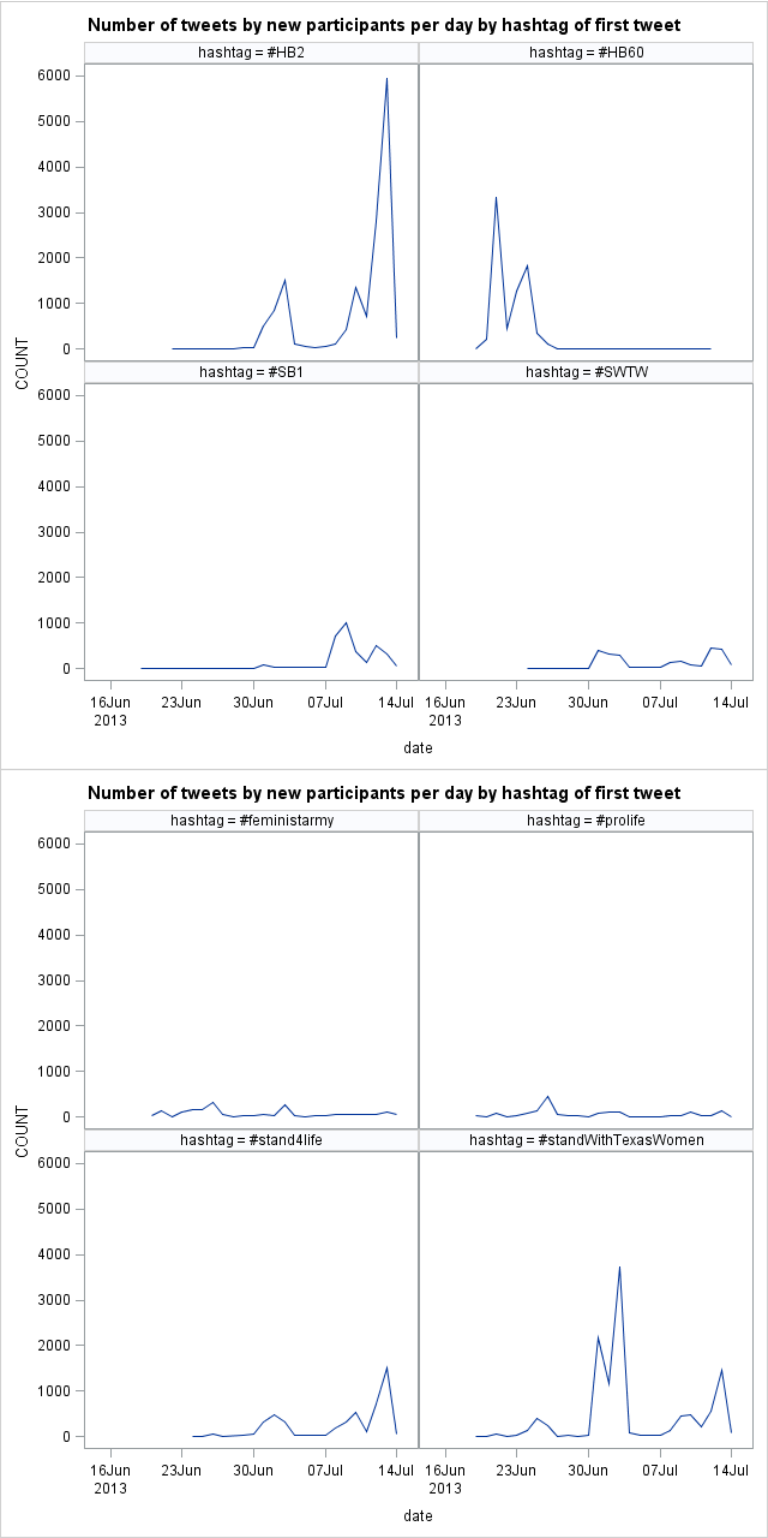


Table 1. Diagnostic statistics for location identification for a sample of 84,125 users with any location data thus far

	Social location estimate	GPS most frequent location	Text or phone lat/long location
Social network location estimate	842/842		
% agree	100%		
GPS most frequent location	29/59	1587/1587	
% agree	49%	100%	
Text location or phone lat/long	107/322	406/869	84,125/84,125
% agree	33%	47%	100%

*Phone lat/long and text location are saved in the same location, so they never both occur

References

Bumpass LL. (1997) The measurement of public opinion on abortion: the effects of survey design. Fam. Plann. Perspect.; 29:177-80.

Davis Jr, CA; Pappa, G. L.; de Oliveira, D. R. R.; and de L Arcanjo, F. (2011) Inferring the location of twitter messages based on user relationships. Transactions in GIS 15(6):735–751.

Gatrell, AC; Bailey, TC; Diggle, PJ; Rowlingson, BS. (1996) Spatial point pattern analysis and its application in geographical epidemiology. Transactions in the Institute of British Geographers. 21(1):265-274.

Steeg, G. V. & Galstyan, A. (2012). Statistical tests for contagion in observational social network studies. CoRR, abs/1211.4889.

Pew Foundation, 65% of online adults use social networking sites (2011); http://bit.ly/OWHYwA.

Appendix A. Outline of process implemented by software written for this analysis

1. Use list of tweet URLs to retrieve profile data and geolocation data for every tweet (TW)
    a. Grab all tweet ids from URLs of tweets
    b. Using tweet ids, get user object from Twitter API
       use lookupUsers([vector],...)
       See: http://cran.r-project.org/web/packages/twitteR/twitteR.pdf
    c. Remove duplicate user ids
    d. Save the following components of the user class
        i. User id
        ii. GPS coordinates
        iii. Cell phone location
        iv. Text string location
        v. Mutual Follower (FT vector)
            1. id
            2. location
        vi. Follower location's of Mutual Followers (FFT vector)
            1. id
            2. location

    NOTE: We find geolocation for everybody using the following method, even those with GPS enabled. This will allow us to test the method in step 8.

2. Find mutual followers of the users who generated TW (call the set of these users T)
    a. Retrieve follower id data from API to generate followers of T (FT)
    b. Retrieve follower id data from API to generate followers of FT (FFT)
    c. Intersection of FFT and FT is mutual followers of T (S)

3. Find mutual followers of S
    a. Find followers of S (FS) (we already have this, it's a subset of FT)
    b. Find followers of FS (FFS) (we already have this, it's a subset of FFT)
    c. Intersection of FS and FFS is mutual followers of S (R)

4. Find geolocations based on GPS where available
    a. For every element of the union of T, S, and R, find out if it has GPS enabled
    b. If it does then get last 100 tweets (use timelines(...) in R)
    c. Find city (county?) most frequently in
    d. Use this city or county as location as BESTGPS

5. Assign geolocations to the union of T, S and R
    a. Use BESTGPS if available, else
    b. Use Geoid if available (from profile), else
    c. Use text location
        i. If using text location, use gazetteer or google API to find city or county
        ii. If gazetteer does not yield a location, then leave geolocation empty

6. For all elements of T (both those where geolocation was assigned using BESTGPS and those without it), find friendship network location (FLOC) using friendship voting.
    a. Find locations of elements of S without GPS enabled using friendship voting among their mutual followers (R) with GPS enabled who had m or more friends with GPS enabled

b. Find locations of elements of T without GPS enabled using friendship voting among their mutual followers (S) with GPS enabled or who had n or more friends with GPS enabled

7. Vary level of m and n to find acceptable match up between BESTGPS and FLOC