In this three-phase assignment, you will do HCI research that explores two questions that nobody knows yet. Only with our powers combined, will we discover the answers at the end of the assignment.

The first question is related to the recent readings, but investigating: how do various factors (features) of *private messages* correlate with your social relationships? Most of the previous work has focused on public social activity, but we spend a lot more time on private messages, and this may be more impactful. The second question is a methodological question: can we as people familiar with our own data, generate more sophisticated and relevant features than the standard 'pile of simple features' approach? Basically, is there any value to hand-curated feature generation?

1   Phase 1: Data Extraction and Labeling

Complete this phase by the check-in date, March 1.

You will use a tool developed by the HCI research group to extract your own historical private social messaging data into a standard format, including text messages, Facebook messages, Twitter direct messages, WhatsApp, Instagram, Kik, etc. Only you will have access to your own data. Download the data extractor from `http://sochiatrist.cs.brown.edu/`, and follow the instructions here: `https://goo.gl/jYAhMG`. Do not anonymize the data or filter by date, as you want all the original data. If you need help with this step, please ask Nedi or Sachin to help with the script. You will need a Mac for this part, so ask Nedi if you don't have one.

Then, randomly sample messages until you get 20 unique people (note that you should randomly sample the messages, not the people, so that you get a stratified sample of people). For each of those people, download the template (`https://goo.gl/BxCHF3`) and enter in the answers to these five questions (based on Gilbert et al., 2009):

1. How strong is your relationship with this person? (-3 = barely know them, 3 = we are very close)

2. How would you feel asking this friend to loan you $100 or more? (-3 = would never ask, 3 = very comfortable)

3. How helpful would this person be if you were looking for a job? (-3 = no help at all, 3 = very helpful)

4. How upset would you be if this person blocked you? (-3 = not upset at all, 3 = very upset)

5. If you starting solely using a different messaging service, how important would it be to add this friend? (-3 = would not matter, 3 = must bring them)

Next, spend a decent amount of time looking through some of your data. You can put it in a spreadsheet and sort by the person you are talking to if it helps. Basically, really get a good

sense of what the data looks like. Come up with 5 important features sorted by how much you believe each would indicate a strong relationship with that person. For example, a feature could be how many unique days that person contacted you first in the most recent 30 days, or a measure of long late night conversations with that person.

2   Phase 2: Programming the Features

Complete this part by March 8 so we can collate everyone's code together.

Next, you will write the code for your five features, using the python template at `http://cs.brown.edu/courses/csci2300/social_analysis.py`. This will force you to be specific about computing the value of the feature from actual messages. But this is not meant to be a test of python programming, so do not hesitate to talk to Nedi or other people in the class for help in transforming your feature ideas into code; Nedi has been allowed to be very helpful here. Your code will need to run on everyone else's computer, so try not to use external libraries and write code that's compatible with *python3*.

If you are a fast programmer, this is also a good chance to install or set up the stats program called SPSS (Brown has a license) so you can run your regression statistics in the next phase.

3   Phase 3: Analysis

Due March 15.

Finally, you will get to run your features on your code to get a value for each feature for each person in your data. Take all of these feature values for the 20 people you have answered questions for, and enter them into the statistical software package SPSS. Basically, each row will be one person and the columns will be the feature values and the 5 responses to the questions.

Then you will run a multiple linear regression (in SPSS this will just be Linear under Regression, but feel free to try out other regression models). If you'd really like to use R instead, feel free, but SPSS has a lower learning curve. You will get results about how yours and everyone's features performed in predicting the strength of your relationships, as well as compared to some simple features previously used in prior literature. Use a similar procedure to the Gilbert paper.

Write a short report about what you did, and what the results were. Bullet points are fine, but be precise about how you talk about the results. Consider how to appropriately write about the explanatory power of the individuals features, and in general how well your model did for predicting relationship strength. At the end of it all, fill out this short form (–TBD on March 8–) with your data so we can see the results as a group.

Your grade will be based on your script for programming your features (2 points), how well your features predict your strong and weak ties (4 points), your multiple linear regression analysis in SPSS or R (2 points), the descriptiveness and correctness of your report (8 points), your check-in

progress (2 points), and running the other students' features on your data (2 points).