

Homework 2

Data Mining, CS 277, Winter 2010

Due Date: in class, Thursday January 21st

Class Web Page

<http://www.ics.uci.edu/~smyth/courses/cs277/>

This homework will focus on a collection of 5 data sets, consisting of Facebook networks and related information collected at 5 US universities. A description of the data, the data itself, and a link to a more detailed technical report are available online from the class Web page.

You are to use MATLAB for this assignment. If you would prefer to use a package such as R or Python instead, please contact me to request an exception—but please note that I cannot check that these software environments have all of the library functions you may need for completing the assignment, etc., i.e., it will be your responsibility to ensure this.

The assignment will require you to submit both a written report (in class) and some MATLAB code (online via EEE).

Background Reading

Download and read the technical report on the Web page (by Traud et al., 2009). Don't worry if you don't understand all the terminology in Sections 2 and 4 in particular, just read and get what you can from the article. You should however pay specific attention to understanding the data itself: how it was collected, what the variables mean, and so forth (e.g., in Section 3 of the paper). The goal of this assignment is for you to see how much of an understanding you can get about this data using all the information available to you. Note that there may be some questions that are unanswered in the paper, e.g., precise definitions of what some variables mean, or how missing values are represented.

You can use any other additional source of information you can find on the Web about this data in your assignment, if you can find any additional information (and please cite your sources). However, since we have a large class and to keep things fair among all students, I would ask that you not email the authors of the article with questions, nor that you email Arthur Asuncion with questions about the data (you can email Arthur later in the quarter with questions if you decide to use this data set for your project).

Write a brief introductory section to your written report that describes how you think the data was collected, what each of the attributes mean, and also a critique of whether or not you think the authors did a good job of describing the data set in their paper. For example, is there information that they should have included that would have helped other researchers who will work with this data in the future or researchers who want to better understand their results?

Downloading the Data

Download the data and check that you can load each of the data files into MATLAB. There should be 5 data files, one for each university. You should verify that you have 2 arrays of data for each university:

- An $N \times N$ binary array where N is the number of individuals in the network. This is represented in MATLAB's sparse format since most of the entries are zero. Basically this means that the locations of all of the 1's in the array are stored as a list of (row,column) pairs and the zeros ignored (which saves a lot of memory). You should learn a little about working with sparse matrices in MATLAB, e.g., type "help sparse." You can operate on a sparse matrix using the usual MATLAB operations like sum, max, min, mean, etc: but note that the results will be returned in sparse format. For example:

```
size(A) % tells us how large A is, N x N
% now compute how many other nodes each node is connected to
s = sum(A); % returns a vector s of size N x 1, which MATLAB will store as a sparse array
s = full( sum(A) ); % returns a "dense" or "full" version of s
```

Note that for the larger data sets you probably won't be able to convert A directly to a dense matrix, i.e., if you type `B = full(A)` you will probably run out of memory (on the larger networks).

- The second array is an $N \times 8$ array of individual node attributes (see the Web page for their names). This array is not sparse, i.e., it is in the more usual "dense" format.

You may want to use MATLAB functions such as `unique.m`, `plot.m`, `boxplot.m`, `mean.m`, `median.m`, etc., below, so you will want to read the documentation on these functions and be able to use them on the arrays above.

Generating Summary Statistics

Do the following for each of the data sets (each university):

Create an additional 9th attribute called "degree", which is the number of other nodes each node is connected to. You should create a vector of size $N \times 1$ from the A array and append it as a 9th column to the $N \times 8$ array of attribute information.

Write a MATLAB function called `summarize.m` that takes as input an array of size $N \times d$ (you will have $d = 9$) and creates the table below. The class Web page provides a template that you can use as a starting point.

Generate a table where the rows correspond to the 9 attributes and where the columns correspond to the following summary statistics for each attribute:

- Number of unique values
- Median value
- Mean value

- Standard deviation
- Most common value (if there is a tie among multiple values, insert the string “tie” in your table, or leave it blank)

In addition, add one more column for an additional summary statistic of your choosing that you think might be of interest - clearly define how this is computed in your report.

Note that since all of the attributes have been encoded as integers, some of the summary statistics (e.g., most common value of ID) will not necessarily make much sense for some of the attributes - but go ahead and generate the table anyway.

Print out the table and hand it in as part of your written report. Submit the code for `summarize.m` online via EEE (under the homework 2 dropbox).

In addition to the table, in your written report, write a few paragraphs about what you observe looking across the 5 tables of summary statistics, e.g., what appears to be common across the data sets and what appears to be different (and why you think this is). What important information about the data might not be revealed in these tables?

Missing Data

See if you can figure out, or guess, at what values are being used to encode missing data for each attribute. Then regenerate the tables above for each of the 5 universities where you remove the values of an attribute that are missing (if there is no missing value for some attribute, you will just get the same numbers as before). One way to do this is to write another version of `summarize.m` that is provided the numeric value (if any) that codes for “missing” for each attribute, and only computes the statistics on the remaining attribute values.

Add a section to your report that shows the new table and that also discusses the differences between the tables without the missing values and those that included the missing values.

Histograms

For each of the 5 universities do the following:

Generate a histogram for the “class year” attribute. Think about how many bins and what bins you want to use for the histogram. (Type `help hist` to find out more about your options for the histogram function in MATLAB). The histogram you produce should be visually informative, e.g., imagine that your histogram will be part of an important report for the CEO of Facebook.

You need to also decide whether you want to include or exclude missing values (if you think there are any missing values).

In your written report add a section that shows the histogram and that comments on what the individual histograms tell you. Any significant differences or commonalities across the universities?

Correlations and Dependencies

Find 2 variables that you think should exhibit some dependence and for these 2 variables, for each of the 5 universities, do the following.

Use a scatter plot or a boxplot to illustrate visually if the two variables do in fact appear to depend on each other. Also compute the linear correlation coefficient between the variables.

Add a section to your report that displays the scatter or boxplot, for each university, and the linear correlations, and add some paragraphs of discussion, i.e., what can you infer from these plots and numbers?

Degree Distributions

For each of the 5 universities:

Generate a plot that plots a histogram of degrees on a log-log scale, similar to the one we showed in class (with the session lengths for ICS Web sessions). The y-axis can be either counts or probabilities (normalized counts), on a log-scale. The x-axis is the degree of a node, on a log-scale.

Add a section to your report that shows your log-log plots, and that discusses whether there is any evidence that the degree distributions obey a power-law.

Optional (no credit, but may be fun to think about): If they don't obey a power-law can you find any other relatively simple distributional model that fits better?

Optional Item

If you have completed all of the above items, feel free to generate one final section for your report where you discuss (with appropriate plots) one other aspect of this data that you find interesting and informative. This is entirely optional and won't be graded, but some of you might find this fun to do. Note that you should not spend time on this unless you have fully completed all of the other required items in the report.