Info 290T Data Mining: *Project Proposal*

# Characteristics of Influencers in Social Networks

## Team Members

Our team is comprised of three undergraduates in the EECS department:


| Member | Specializations | Contact |
|---|---|---|
| David Wu | Statistics, full-stack Web development | dwu7401@gmail.com |
| Jimmy Wu | CS theory, full-stack Web development | j.wu@berkeley.edu |
| Alton Zheng | UX, mobile, full-stack Web development | altmnop@gmail.com |


## Analysis Objective

We propose a general analysis of social graphs to extract the features that determine influence in social networks. Let the dataset consist of a graph $G = (V, E)$, either directed or undirected, representing users and relationships between them. Furthermore, let each node/user $v$ be associated with a feature vector $\langle f_1, f_2, \ldots, f_m \rangle$ describing that user. Our analysis is in two steps:

1. Generate, for each $v$, a nonnegative value representing its *influence* in $G$. There are many algorithms suited to this purpose: PageRank, degree centrality, eigenvector centrality, etc.
2. Learn a model correlating feature values with influence rankings—that is, a function that approximates the values in step (1).

This model is the final output. It can then be used, among other things, to:

- Alert the managers of the social network to some of the qualities that make for valuable users—what do our most valuable users' profiles look like? Subsequently, this good behavior can be encouraged in hopes of fostering a graphically rich social community.
- Predict the eventual influence of a new node (that is, a new user in the social network) before its surrounding graph topology has been established.

## Choice of Dataset

We have found two promising datasets:

- A Google+ dataset from the Stanford Large Network Dataset Collection[1]
- The Yelp Dataset Challenge[2]

The Google+ dataset has a well-connected, dense graph structure, but only six user features. The Yelp dataset, on the other hand, contains a significantly sparser graph, but much richer user dimensions. We plan to first apply the method to the Yelp dataset; if we find that its graph topology or other factors make its usefulness unsatisfactory for our purposes, we will then pursue results on the Google+ data.

## Testing Plan

One nice thing about the two-stage method is that once influence values have been assigned to nodes, the graph structure no longer matters. Thus, we can perform cross-validation by setting aside different subsets of the vertex set for testing.

A more subtle aspect, however, is which influence metric to use. Since the choice of ranking algorithm directly influences the final model, we will need to try several kinds, then train and test the model using each.

## Schedule

| Week | Tasks | Notes |
|---|---|---|
| April 10 - 16 | Parse & clean data, build graph & user modules | |
| April 17 - 23 | Learn on Yelp data, test, assess output | |
| April 23 - 30 | Learn on Google+ data, test, assess output | Code & data due April 30 |
| May 1 - 7 | Prepare presentation | Presentation due May 8 |
| May 8 - 14 | Prepare paper | Paper due May 14 |

[1] `https://snap.stanford.edu/data/egonets-Gplus.html`
[2] `http://www.yelp.com/dataset_challenge`