University of California, Berkeley

Yelp Elites in the Wild

David Wu
dwu7401@gmail.com

Jimmy Wu
j.wu@berkeley.edu

Alton Zheng
altmnop@gmail.com

May 14, 2014

1 Introduction

Social media is an ever-growing subfield of software and arguably the most important development in the past decade of Internet applications. As people use networking sites such as LinkedIn and Facebook, as well as utility sites like Yelp and Quora, companies are able to extract fascinating insights into users' habits. However, these reservoirs of social knowledge can only live up to their hype if they are able to engage an active, involved user base—in other words, the quality of a social network matters. One way to gauge and improve the quality of a service is to identify its most valuable members and study their characteristics. From there, a company managing the social network can design its products so as to promote engagement and build a richer community. In this paper, we explore strategies to do just that: find outstanding social community members and discover what makes them valuable.

This project focuses on data from the Yelp Dataset Challenge [1]. The dataset is abundant in mineable information, encompassing over 15,000 businesses with attributes; a 70,000-user, 150,000-edge graph; 335,000 reviews, and more.

Our work is open-source and can be viewed at `https://github.com/jimmywu126/yelp_influence`.

2 Problem

As alluded to earlier, we hope to identify valuable members of the Yelp community. With the given dataset, we are unable to discover prima facie what makes individuals popular and valuable to the social network, and as a result, are unable to easily enumerate what makes for a rich social graph.

3 Solution

At first glance, one might turn to the user graph itself to identify top social citizens—say, by computing some influence metric for each user and learning what makes someone influential (or not). Indeed, this was our original approach: using Google's famous PageRank algorithm as a proxy for a user's importance, we performed regression analysis to correlate user attributes with PageRank. However, the results were relatively poor, with no particular user feature telling us much about his/her PageRank.

Taking a step back and exploring the graph topology, we discovered the reason why: the Yelp social graph is extremely sparse. It contains $|E| = 151,516$ edges and $|V| = 70,817$ nodes, giving it a density (defined as the fraction of the $\binom{|V|}{2}$ possible edges that actually exist) of about 6.04×10^{-5}. Worse yet, the distribution of node degrees (i.e. the number of friends a user has) seems to follow a power law, with over half of users friendless (see figure 3.1). Overall, the metrics clearly indicate that the service is not distinctly focused on the social aspect of user interaction. For our purposes, this means that PageRank is not a sufficient label, at least in itself, on which to base our notion of user quality.

Our solution is to use an easily-overlooked user attribute: the Yelp Elite tag. Yelp labels certain users as particularly valuable by designating them as 'Elites'; these users are identified by the company's internal Community Managers on an individual basis; however,

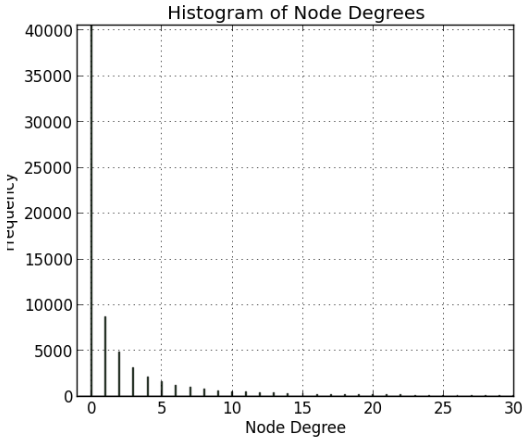

Figure 3.1: Distribution of users' friend counts

Yelp publishes no specific metrics or rubric for granting Elite status. Yelp simply states that it looks for 'authenticity', 'contribution', and 'connection.' These relatively vague metrics indicate the possibility that the criteria for Elite membership vary between different community managers. But the good news is that they are hand-picked, and thus make good candidates for top users. Our new approach, therefore, attempts to learn the most crucial factors that contribute to Elite membership. We utilize several different machine learning paradigms, including linear regression, SVM, and Bayesian classifiers, to tackle this problem, and we make use of the plethora of raw data on users, relationships, and reviews provided by the dataset.

4 Details

4.1 Data and Tools

We begin with the 'users' file, a subset of the Yelp Dataset Challenge dataset. Each line contains a user dictionary, organized as follows:

```
{
  'user_id': (encrypted user ID),
  'name': (first name),
  'review_count': (review count),
  'average_stars': (float average, e.g.
  4.31),
  'votes': {(vote type): (count)},
  'friends': [(friend user IDs)],
  'elite': [(years elite)],
  'yelping_since': (date, e.g. '2012-03'),
  'compliments': {
  (compliment_type): (num compliments),
  ...
  },
  'fans': (num fans)
}
```

Later we incorporate the 'reviews' file, whose objects are as follows:

```
{
  'business_id': (encrypted business ID),
  'user_id': (encrypted user ID),
  'stars': (star rating),
  'text': (review text),
  'date': (date, e.g. '2012-03-14'),
  'votes': {(vote type): (count)}
}
```

To perform our analysis, we use several Python libraries: `NetworkX` for graph manipulation, `numpy` for general scientific computing, `matplotlib.pyplot` for visualization, `scikit-learn` for machine learning, and `NLTK` for natural language processing.

4.2 Linear Regression

Our first attempt at learning weights on user features is an obvious choice: linear regression minimizing sum-of-squares error. Using feature vector $f = \langle$review count, average rating, friend count, length of Yelp membership$\rangle$, we trained a model in the form of a weight vector w. Unfortunately, when applied to test data, the model achieved an R^2 score of only 50%.

We subsequently bolstered our feature set, adding vote counts (when other users rate your reviews as 'cool', 'funny', or 'useful') to the feature vector and refining membership length by using months rather than years. Lastly, we replaced friend count with PageRank score; at the cost of significantly greater computational complexity, PageRank encoded the graph topology more precisely than the simple local node degree [2]. The result was a modest improvement to 56% R^2 score (varying depending on our randomized hold-out test points).

Despite unimpressive results, however, the regression model managed to reveal the traits that mattered most. Here is the weight vector w learned:

```
months_member: 0.453727926216
review_count: 20.4070788195
cool_vote_count: -18.7091084595
pagerank: 7.01379346803
```

```
funny_vote_count:  5.66321051754
average_stars:  0.0172153215606
useful_vote_count:  8.8595850916
```

4.3 SVM

Using the more sophisticated technology of support vector machines (SVMs), we again trained a model—a separating hyperplane (to be specific, an SVR, support vector regression model, but the underlying mathematics is essentially the same). We did this with linear, quadratic, and cubic kernels, but the results were disappointing: no SVM scored significantly better than 40% on the R^2 test. At the same time, these models were time-consuming to train.

At this point, the failure of both linear regression and SVMs with various polynomial kernels to produce satisfactory results forced us to consider that the underlying structure of Elite status is very nonlinear; after all, the label is awarded manually by real people, via non-computational means, so there is no reason to believe it should be so. We then turn, logically, to probabilistic techniques.

4.4 Naive Bayes

Our last tactic was the naive Bayes classifier, a bayesian network that assumes independence between all features. Though simple in theory, naive Bayes turned out to be our most accurate model.

The classic naive Bayes classifier requires discrete feature values, so that they can be mapped to probabilities. Since most of our features are continuous or discrete but over large domains (PageRank, months member, etc.), we leveraged the Gaussian naive Bayes model. It assumes all distributions $P(X|y)$ for features X and labels y are Gaussian—an aggressive simplification, but a reasonable one in practice for continuous variables.

The result: 90% classification accuracy across our randomized tests. There is one pesky detail, however: since Yelp's users are only 7% Elite, the model favored guessing 'not Elite'—that is, recall hovered around 70% for Elite members. By using stratified sampling of our data points, we forced Bayes to generalize better, boosting recall up to 80-85% while maintaining the overall accuracy rate.

As a final improvement, we began to integrate Yelp's review data for these users into the training process. Inserting a user field for 'average review length', accuracy improved to 91%, and recall to 82-87%. These gains are modest, but with further extraction of meaning from text using NLP methods, we believe the classifier can be improved even further.

5 Related Work

This topic has been studied extensively by researchers in the information sciences as well as the social sciences. A closely related part of the literature discusses the problem of detecting sybils, which are fake accounts on social networks whose solve purpose is to spam legitimate users, or various other undesired activities. A sybil can be seen as the polar opposite of a strong user, which is the focus of our work. In that sense, by answering this question, we are also exploring methods of detecting sybils. An interesting paper written by researchers from Peking University and UC Santa Barbara, leveraging data from LinkedIn and the Chinese social network Renren, seeks to solve this precise problem [3].

In another similar work, Cornell researchers try to answer the question, "If we can convince a subset of individuals to adopt a new product or innovation, and the goal is to trigger a large cascade of further adoptions, which set of individuals should we target?" [4] This question is very similar to ours; after all, the task largely reduces to finding the nodes in a social network that have the most influence. Yelp Elites wield considerable social influence, and it would be interesting to see what their algorithms predict would be the most influential people in Yelp's social network.

6 Conclusions and Future Work

In our work, we explored techniques for identifying high-value users within a social network. Starting with an attempt at correlating attributes with PageRank, we realized that the sparsity of the Yelp user graph makes it inconducive to purely graph-theoretic analysis, though graph properties were later useful as one user attribute among many others.

Thus, we quickly shifted our efforts towards learning on Elite status, which we found to be an insightful label. We ultimately discovered that naive Bayes, despite its conceptual simplicity, was the most effective predictor of Yelp Elite status, performing both accurately and swiftly. Other more sophisticated approaches were either too slow, or simply yielded poor results.

One of the main areas that we believe deserve to be pursued more closely is business review text, which may help determine the quality of the users who write them. While we incorporated many features of users, the most valuable corpus in the Yelp dataset is clearly its human-written reviews. We believe it makes a great deal of sense to use this data as an additional predictor of Yelp Elite status. Currently, we include the user's average review word count as a feature, leading to modest gains. As of this writing, we are working to incorporate the reading level (the level of sophistication of a person's writing), a computationally intensive task.

In the future, we wish to explore the possibility of integrating other interesting indicators, such as unique word count or number of spelling mistakes. Alternatively, we can make the subject finer grained, and try to predict the quality of individual reviews; ideally, this would involve a manually labeled set of 'good' reviews for training, which is currently not available.

In all, our approaches are a good start towards improving social network quality by studying their most valuable members. The possibilities are seemingly endless, and it is our belief that, as the business value of social information continues to rise, researchers will turn their attention to designing and encouraging more robust social graphs.

References

[1] Yelp Dataset Challenge. `http://www.yelp.com/dataset_challenge`

[2] S. Brin and L. Page. "The PageRank Citation Ranking: Bringing Order to the Web." `http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf`

[3] Yang et al. "Uncovering Social Network Sybils in the Wild." `http://arxiv.org/pdf/1106.5321.pdf`

[4] Kempe et al. "Maximizing the Spread of Influence through a Social Network." `http://www.cs.cornell.edu/home/kleinber/kdd03-inf.pdf`