Chang Liu
Richard Kim
Joe Palin

yelp207@gmail.com
AM207 Spring2015

WHO IS THE MOST INFLUENTIAL ONE ON YELP?

Introduction

To promote 'the word of mouth' over social networks in today's digital economy, Yelp's business model relies on its users to generate thousands of thousands of reviews on various types of businesses everyday. As a part of it's marketing effort, Yelp selects a limited number of users to be 'Elite Yelpers' and provides special incentives (i.e. free perks such as passes to exclusive events at new restaurants and bars). Its aim is to encourage the 'Elite Yelpers' to write high-quality reviews that will influence and attract other users such that information about the new business 'cascades' down its network. Consequently, Yelp would be interested in discovering among the users who are able to engage and influence the largest number of other users. The question then is, Who should they be? This is exactly what we aim to find out in the Influence Maximization Problem.

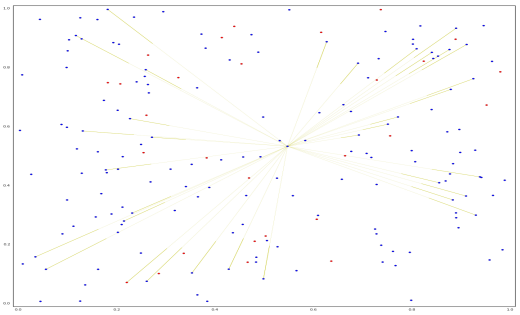

Figure 1, Independent Cascade of one user's influence

Data

The Yelp Challenge Dataset releases:
1.6M reviews by 366K users for 61K businesses;
481K business profiles and social network of 366K users.

We visualize one independent cascade of the influence of a random user over its neighbors after 3 time steps. The blue dots represent activated nodes with action attribute 1; the reds are nodes that have not been influenced and have action attribute 0.

Approach

We model the Influence Maximization Problem as a graph. A node in the graph represents a user, and it has action attribute '1' or '0', which indicates whether or not user is 'activated' (i.e. has reviewed a particular business) and a directed edge (A, B) represents: user A has some influence on user B to activate (i.e. review the same business). We assume that if both user A and user B wrote reviews on n businesses and user A wrote k>0 reviews prior to user B, then user A has k/n probability of influencing user B.
The goal of Influence Maximization Problem is to find the initial set of nodes that has maximum influence over the network after some time steps t.

The problem is known to be NP-Hard and thus prohibitively expensive to find the actual solution as the graph becomes larger. We design and compare 4 approximation algorithms: Greedy Algorithm (with memoization), Genetic Algorithm, Simulated Annealing, and randomization.

Conclusion

We are currently examining the outcomes of the four different approximation algorithms and will continue to experiment with different sizes of initial set of users.

Citations & Links
Economics and Computation, D. C. Parkes and S. Seuken, Cambridge University Press 2016.
http://www.yelp.com/dataset_challenge