Connecting the World:

Building a social network recommendation system for both local and global connectedness

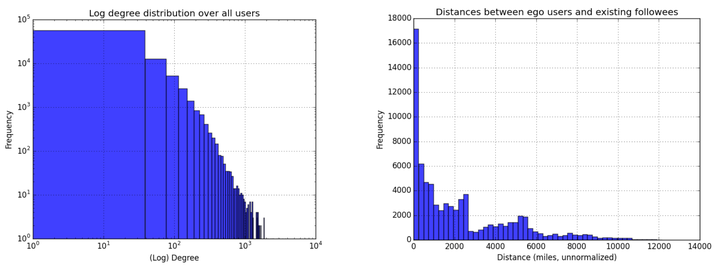
Abraham Botros
abotros@stanford.edu

Problem

- We all share the same home – this pale blue dot we call Earth.
- Yet we allow ourselves to be thoroughly divided by things such as cultures, skin color, and invisible lines on the ground.
- Can we encourage new understandings, relationships, and perspectives in our online social networks?
- Previous approaches:
  - Recommendation based on interests, popularity, network structure (YouTube[1], Twitter[2])
  - Geographic routing, relation to friendship probability (LiveJournal[3])
- Goal: Build social network recommendation system that promotes diverse, distant, and yet relevant relationships between users across the globe, in addition to familiar nearby users.
  - Recommend numerous similar users with strong network importance that are nearby in location
  - Recommend a few similar users with strong network importance but that are far away in location

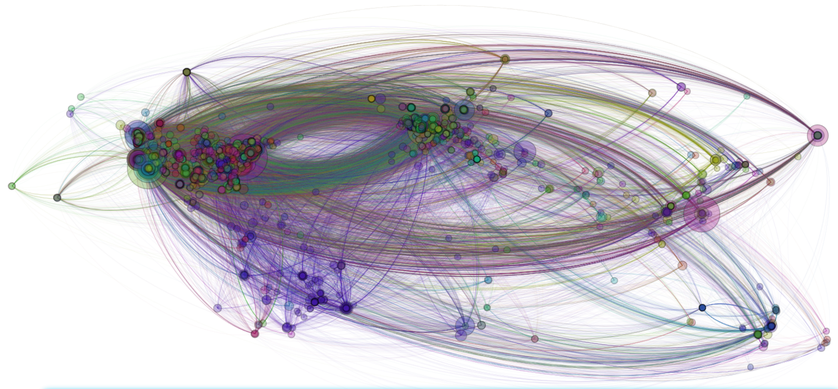
Data

- Twitter SNAP ego network dataset; 81K nodes, 1.8M edges, 973 egos. 81306 WCC/68413 SCC. 0.5653 clustering.
- Filter locations: 57K nodes, 1M edges, 746 egos
- Node ID :: Twitter API :: Profile location :: Google Maps Geocoding API :: (lat, lng) for each user



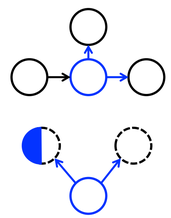
(Left) Log degree distribution over all users shows high skew; the vast majority of users have relatively low degree (less than 500), with a few users with extremely high degree (2000-4000). (Right) Histogram of unnormalized distances (in miles) between ego users and their existing followees; we see most ego user followees are relatively close to the ego user.

References, footnotes: 1. Davidson et al. YouTube recommendation. RecSys '10.
2. Hannon et al. Twittomender. RecSys '10. Liben-Nowell et al. Geographic routing. NAS '05.
4. (±2.5, 2.0, 0.5, 1.0, 1.0, 1.0, 1.0, -1.0, -1.0)



Features

- **Geographic distance**
  - Distance from ego to candidate using Haversine formula
- **Network properties**
  - In-degree, out-degree, degree
  - Degree centrality, closeness centrality, PageRank
- **Topic-interest generation**
  - Intractable/impractical to deduce from accessible Twitter data
  - Generate randomly for egos, propagate along edges



- For each ego user *u*
  - For each non-followee *v* (sample 2000/57000 for tractability)
    - Distance $D_{u,v}$
    - Network features $F_{net,v}$
    - L2-norm, interests, $\|I_v, I_u\|$
    - Average L2-norm, interests, $\overline{\|I_v, I_{\text{followees of } u}\|}$

$$F_{u,v} = \left[D_{u,v}, F_{\text{net1},v}, F_{\text{net2},v}, F_{\text{net3},v}, F_{\text{net4},v}, F_{\text{net5},v}, F_{\text{net6},v}, \|I_v, I_u\|, \overline{\|I_v, I_{\text{followees of } u}\|}\right]$$

Scoring and Recommendation

Weights[4]: $W = \left[w_D, w_{N1}, w_{N2}, w_{N3}, w_{N4}, w_{N5}, w_{N6}, w_{I,\text{direct}}, w_{I,\text{avg}}\right]$

Score: $S_{u,v} = W \cdot F_{u,v}$

- **2 modes: Distance-penalizing, and Distance-rewarding**
  - w_D distance weight set to large **negative** value to **penalize** large distances
  - Set to large **positive** value to **reward** large distances
- Sort lists according to scores; take top 16 distance-penalizing, top 4 distance-rewarding, for a total of **20 recommendations per ego user**

Results

(All results are averages over all recommendations for all ego users. All values are normalized to the [0,1] range based on observed mins and maxs for their variable over all non-followees for all ego users.)

Previous followee distance	**Distance, penalized**	**Distance, rewarded**	Number "long-distance" recommendations
0.27493	0.07199	0.79342	4.24531

The location-influenced recommendation was successful! When *penalizing* distances from the ego to recommended users, we get an average normalized distance of 0.07199 (quite low; corresponds to ~1000 miles). When *rewarding* larger distances, we indeed get a much larger average normalized distance of 0.79342 (quite high; corresponds to about 10000 miles). Lastly, if we quality a "long-distance" recommendation as one that is farther than the average previous followee distance, we see we get the exact number of "long-distance" recommendations we desired. Everything here worked as desired!

In-degree	**Out-degree**	**Degree**	**Degree centrality**	**Closeness centrality**	**PageRank**
0.00619	0.01943	0.01180	0.00987	0.37263	0.00107

Since these values are normalized, the equivalent unnormalized degree for the normalized 0.01180 is approximately 45, which is actually near the middle-to-upper-end of the bulk of the degree distribution. In addition, since we are sampling only 2000 of ~57000 candidates for each ego user, we expect that the upper end of what we encounter (and could possibly even recommend) might be users with degrees of around 40-50. So overall, good results, too!

Topic-interest L2-norm, ego	**Topic-interest L2-norm, ego followees**
0.23345	0.54832

Recalling that smaller L2-norms between topic-interest vectors correlate to higher similarity, we see we do relatively well on average for recommending similar new users to an ego user based on topic-interest vectors (0.23345 on the [0,1] range). We do not do quite as well for recommending similar-interest users compared to the ego's followees, but we expect larger variance here anyway due to further averaging and the pseudo-random method for interest generation for ego followees.

Conclusion

- Project overall successful! Results exactly what we expected. We are able to recommend similar users with strong network features while also either penalizing or rewarding locations at will.
- Future work would involve digging deeper into centrality network features and degree redundancy, penalizing large distances a little less in the distance-penalizing mode, and much more exploration of weight vector settings. Also better metrics for user similarity, and using real NLP-derived interests.