Structural and Dynamic Analysis of Foursquare Network

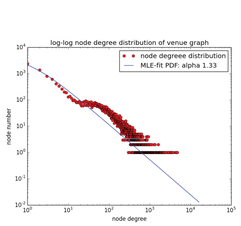
Huimin Li, Yang Zhao, Ningxia Zhang

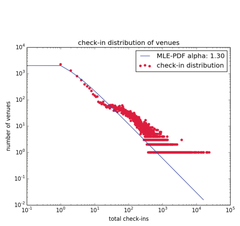
Introduction

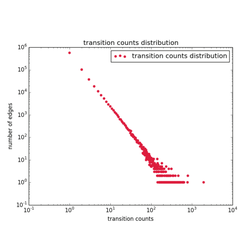
LBS not only adds location context to our social activities, but also presents graph structure consisting of individuals and places in certain relationship. In this project we investigated the interactions between venues in Foursquare network.

Dataset & Basic Analysis

With venue information and transitions from the Foursquare Dataset, we built our transition graphs and conducted initial anaylsis on graph structure.







- Deg. distribution follows power-law, α = 1.33
- “Plateau” might mean mixure with Guassian
- Extremely skewed bow-tie shape
- Transition count follows power-law, α = 4.12 (!)

NO.	Names	Value
(1)	Total Node Number	16218
(2)	Total Edge Number	771831
(3)	The largest SCC	53.10%
(4)	The in-component of the largest SCC	0.16%
(5)	The out-component of the largest SCC	46.65%
(6)	The disconnected components	0.09%

Community Analysis

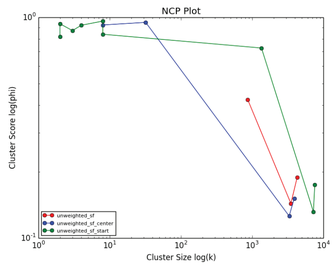
- Modified weighted Girvan-Newman

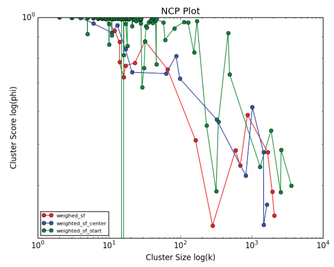
$$\delta_v(v,w) = \frac{\delta_{sv}}{\delta_{sw}}(1+\sum_x \delta_s(w,x))\frac{1}{\sqrt{W(v,w)}}$$

- Unweighted and weighted CNM

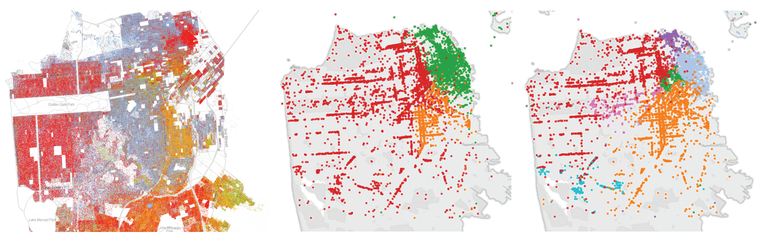
$$Q = \sum_i (e_{ii} - a_i^2) \qquad \Delta Q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j)$$

- Conductance-based overlapping algorithm





Aside from NCP plots, we visualized venues on the map with community color-coded.



- Communities are clustered geographically
- Community map mirrors racial segregation map
- Weighted CNM outputs finer results
- Each community has distinct category distribution, e.g. most common category of red community is Chinese restaurant

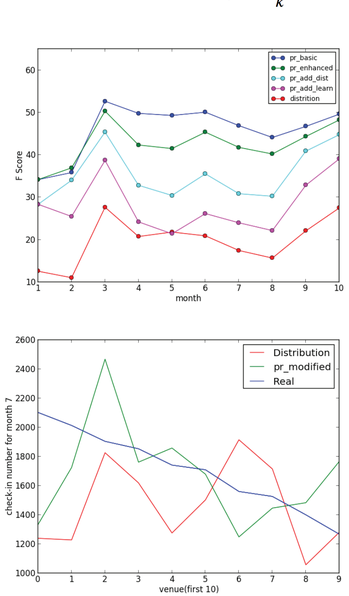
Evolution Analysis

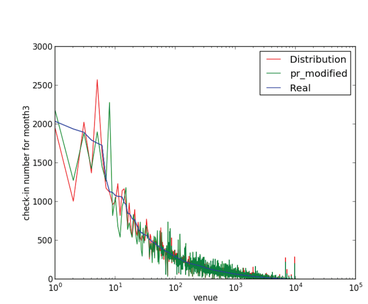
In this section, we used two methods to predict future check-in numbers. The first approach is based on previous check-in distribution in the graph:

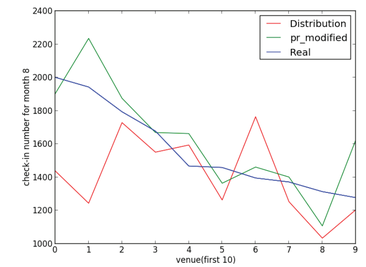
$$c_v^{(t+1)} = M \cdot \frac{c_v^t}{\sum_v c_v^t}$$

We also modeled the evolution using PageRank with multiple modifications, such as enhancing the graph with most recent check-ins and learning from previous predictions.

$$r'_j = \sum_{i \to j} r_j \cdot \frac{(\lambda + dist_j) \cdot f_{ij} \cdot L_j}{\sum_k ((\lambda + dist_k) f_{ik} \cdot L_k)} + \sum_{deadend\ i} \frac{r_i}{N}$$







- Both models mimic the real check-in patterns
- Distribution-based performs better in general
- Modified PageRank models highly visited nodes more accurately