Interaction of Friendship and Geographic Movements in Location-Based Social Networks

Chen Guo (cguo2), Bowei Ma (boweima)

December 12, 2016

1 Introduction

People interact with each other both online and offline. The rapid development of social network have gained us the access to considerable data about online communication or friendship. However, it's much harder to track people's movements in the real world. Location-based social network service offers us a good opportunity to study the interaction between people in cyberspace and real world, based on the checking-ins by users.

Much work has been done to investigate the mobility patterns and temporal effects of human movement. People tend to have daily and weekly periodical behaviors.

What we are most interested in is the relationship between human behavior and location's effects. As we have the online and offline social network at the same time, it's very natural to ask: how does friendship interact with geographical movements.

Lots of study has been done regarding the interaction between social network and geographical network, like social-spatial properties in [4] and check-in behaviors in [5]. We specifically narrowed our goal to modeling the users' behaviors and venues' functions in more detail and the relationship between them. What we've done include:

- Come up with an index to measure the extent to which one user likes to travel around instead of staying in a relatively small area.
- Come up with an index to measure the extent to which one location/venue is likely to bring strangers together instead of maintaining friendship.
- Further study about the relationship of the above two indexes. For example, is it true that people with shorter traveling distance are more likely to visit the places that tend to bring strangers together?

2 Related Work

2.1 Measuring Urban Social Diversity

In [1], researchers focused on the very specific topic of urban social diversity. Understanding how human mobility enhances the social diversity of places is an important question for urban planners and system designers alike.

Based on these two networks (user network and location network), they measured the social diversity via brokerage, serendipity, entropy and homogeneity. They discussed the bridging (bringing together strangers) and bonding (bringing together friends) qualities of different places.

However, their work focused more on the features of locations and their influences to the evolution of urban areas. In addition to that, we want to study the relationship of these metrics to different types of users.

2.2 Friendship and Mobility

In [3], Cho and Myers explored the patterns of friendship and mobility using location-based social network check-in data and cellphone location data. They found that people experience a combination of strong short range spatially and temporally periodic movement that is not impacted by the social network structure, while long-distance travel is more influenced by the social network ties.

Their main task is to build a model to predict the mobility of users. Considering the temporal and geographic periodicity of users' movements. They built a Periodic Mobility Model considering both of these as components. As comparison, they built Periodic & Social Mobility Model as well which took social influences into consideration.

They've done thoroughly analysis of human mobility. However, they treat all the users equally and the results they got are the average behaviors of all the users. But we believe there's divergence between different types of users. We will explore more on the mobility for certain group of people.

2.3 Returners and Explorers

In [2], researchers used Global System for Mobile Communications (GSM) data and GPS data as the digital traces of human mobility. Similarly to [3], they have found that there exists some recurrent pattern of human mobility (i.e. traveling between home and work, visiting the same places a lot). However, these recurrent activities seem to have vanishing influence on the total mobility for some people, which are called explorers compared with returners.

They defined a radius of gyration to characterize the typical distance travelled by an individual, making use of mass center of places. Then they can classify the users into two groups: those whose typical distance traveled can be dominantly defined by the top k places are called k-returners compared with k-explorers.

However, human is more complicated than a two-type group. Most of the time, we can't split a group using some fixed standard. Instead, using some continuous or multi-class index to measure this characteristic might be more useful. Also,although this article discussed a lot about how these two types of people behave in social relations, they hardly mentioned how these two types of people interact with venues and locations in the geographical world, which is what we expect to work on further.

3 Data Sets

As social network becomes more involved in people's lives, an increasingly large amount of data set becomes available, not only about online friendship but also about movements in real world. However, our data set must contain certain information: user friendship network, information about venues including coordinates and optionally categories, and features about check-in activities (like user, location and time).

We considered data sets provided from 4 different location-based social media: `Brightkite`, `Gowalla`, `Foursquare` and `Yelp`. Among these data sets, `Foursquare` and `Yelp` has the categorical information about the venues. However, `Foursquare` API does not provide developers with the access to the data about user friendship. `Yelp` only provides the amount of check-ins per hour for certain venue. Considering all the pros and cons of these available data, we decided to use data set from `Brightkite` and `Gowalla`.

`Brightkite` data set has social network of 58,228 nodes and 214,078 edges, and 4,491,143 check-ins. Since the size is smaller than `Gowalla`, so by now we are working on `Brightkite` data set to build our models and conduct experiments. After we have the framework and models, more test are also done on the larger data set from `Gowalla`.

4 Network Model

A lot of systems in the real world involve multilayer networks and interactions among them. Specifically for our example, we are dealing with both geographical and user friendship networks.

The goal of our project is to understand the different behaviors of users in terms of the extent to which he/she likes to explore new things, the social roles of different venues are playing, and the relationship between these two. Hence, the basic structure of our network model could be summarized using Figure 1.

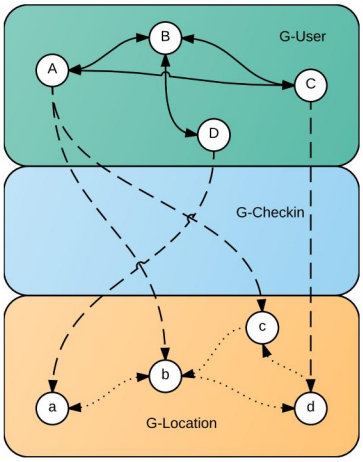


Figure 1: Diagram of $\{G_U, G_L, G_C\}$

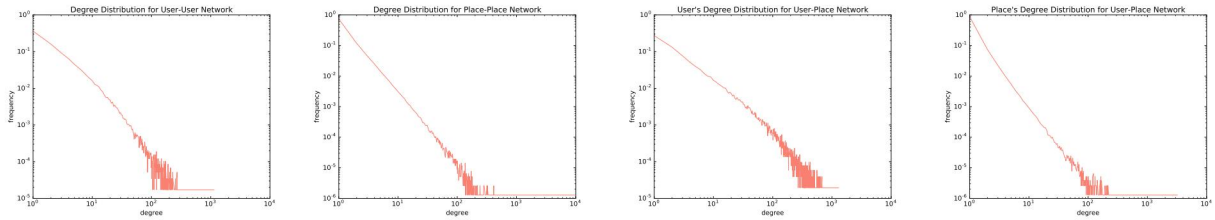
Notation	Meaning
G_U	The graph of user friendship
U_i	The ith user node
G_L	The graph of locations
L_i	The ith location node
P_i	The geopgraphicla position of L_i
G_C	The check-in graph
$d_U(U_i)$	The degree of U_i in G_U
$d_C(U_i)$	The degree of U_i in G_C
D_{ij}	Distance between L_i and L_j
$\text{nbr}_C(U_i)$	The neighbors of U_i in G_C
$w_{(U_i,L_j)}$	Weight of edge (U_i, L_j)

Table 1: Notations

4.1 Framework

Intuitively, we build a model which is similar to the one described in [1]. This three-layer model can be described as a set of simple graphs $\{G_U, G_L, G_C\}$.

In Figure 1, G_U is an undirected graph with nodes as users. G_L is an directed graph with nodes as venues. Theoretically, we are supposed to use real-world roads between different places to build G_L. However, here we just use a simplified way to connect two location nodes as long as one user check in consecutively in these two places. We also model these edges as weighted edges with $w_{(s,t)}$ as the times of transitions from s to t. If we ignore the weights, G_L could also be considered as undirected. G_C is a directed bipartite with users on one side and locations on the other side. In



(a) $d_U(U_i)$ **(b)** $d_L(L_i)$ **(c)** $d_C(U_i)$ **(d)** $d_C(L_i)$

Figure 2: Degree Distributions

G_C, we define all the edges starting from user nodes and ending with location nodes, which from a bipartite graph. These edges are also weighted as $w_{u,l}$ as the times of check-ins from user u at location l.

Table 1 holds the notations that we would use in the future modeling and analysis.

4.2 Summary Statistics

For the data set `Brightkite`, we have the degree distributions for all the three graphs in log-log scale. Notice that for G_C, we want to separately plot the degree distribution for user nodes and location nodes. All the 4 distribution plots are in Figure 2. All of them are showing long-tail patterns like typical real-world social-networks. (Note: we treat geographical network as undirected here.)

We also summarize some statistics here in Table 2, where N_{LWCC} is percent of nodes in LWCC, E_{LWCC} is the percent of edges, D_{LWCC} is the diameter in LWCC and CC is the average clustering coefficient.

Graph	Nodes	Edges	N_{LWCC}	E_{LWCC}	D_{LWCC}	CC
G_U	58,288	214,078	0.9744	0.9947	7.117	0.1723
G_L	772,966	1,549,224	0.9704	0.9845	12.007	0.1514
G_C	824,372	1,076,043	0.9631	0.9786	/	/

Table 2: Summary Statistics for Graphs

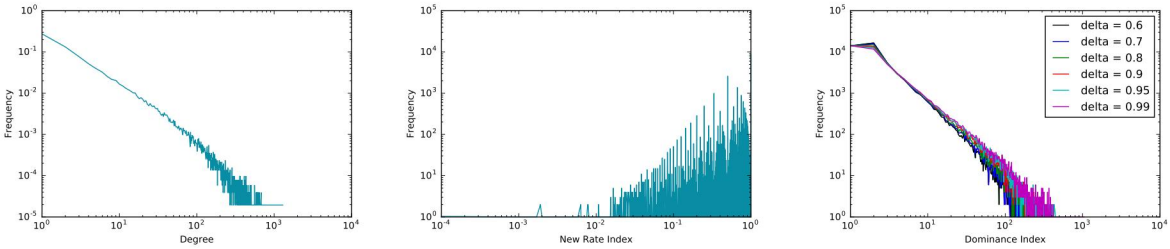
5 Methodologies and Findings

5.1 Modeling for Users

In order to measure the extent to which one user tends to explore different places rather than traveling among some limited amount of places, we define an index E_i as the exploration index of user U_i. The higher E_i is, the more U_i tends to explore unvisited venues.

There are various ways to model E_i from different points of views.

- Most easily, we can use the $d_C(U_i)$ as **Degree Index**.
- From the aspect of new places: the proportion of new places in the second half of check-ins. We can this **New Rate Index**.



(a) Degree Index **(b)** New Rate Index **(c)** Dominance Index

Figure 3: Degree Distributions

- Make use of radius of gyration in [2] to measure the average traveling distance of a user. We define $E_i = \min\{k | r_j^{(k)} > \delta r_j\}$ as **Dominance Index** where

$$r_i = \sqrt{\frac{1}{N}\sum_{j=1}^{d_C(U_i)} w_{(U_i,L_j)}(P_i - P_{cm})^2} \qquad r_i^{(k)} = \sqrt{\frac{1}{N_k}\sum_{j=1}^{N_k} w_{(U_i,L_j)}(P_i - P_{cm}^{(k)})^2}$$

  and P_{cm} is the position of center of mass, $P_{cm}^{(k)}$ is the center of mass from the k most frequently visited places, and δ is a constant. This measure basically tells us that a user's movement is dominantly defined by the first k places he goes to most frequently. Choosing different δ will also yield different performance of the index, the larger δ is, the stricter the measure is.

Figure 3 shows the value distribution of these index (Including different values of δ for **Dominance Index**). In order to understand the correlation of these indexes, we also plot the scatter matrix of these three index in Figure 4. These three indexes all have their limitations. For example, there are lot of users that only check in once, we are not able to model their behaviors very accurately.

5.2 Modeling for Locations

In order to measure the extent to which a venue tends to bring strangers together instead of maintaining friendship, we define the bringing index B_i for location L_i. The higher B_i is, L_i functions more as a hub that people who don't know each other might stop by.

There are also various ways to model B_i from different points of views.

- From the aspect of G_U: $\text{nbr}_C(L_i)$ form a sub-graph in G_U. The reciprocal of average degree of this sub-graph as **1/Degree Index** can measure the friendship of this group of people. The number of edges existing divided by the number of all possible edges is similar, which yielding **Edge Number Index**. Explicitly, we have

$$1/\text{Degree Index} = \frac{1}{\frac{1}{|\text{nbr}_C(L_i)|}\sum_{v\in\text{nbr}_C(L_i)} d_{\text{sub}U}(v)}$$

$$\text{Edge Number Index} = 1 - \frac{\sum_{u,v\in\text{nbr}_C(L_i)} 1_{e_{u,v}}}{|\text{nbr}_C(L_i)|(|\text{nbr}_C(L_i)|-1)/2}$$

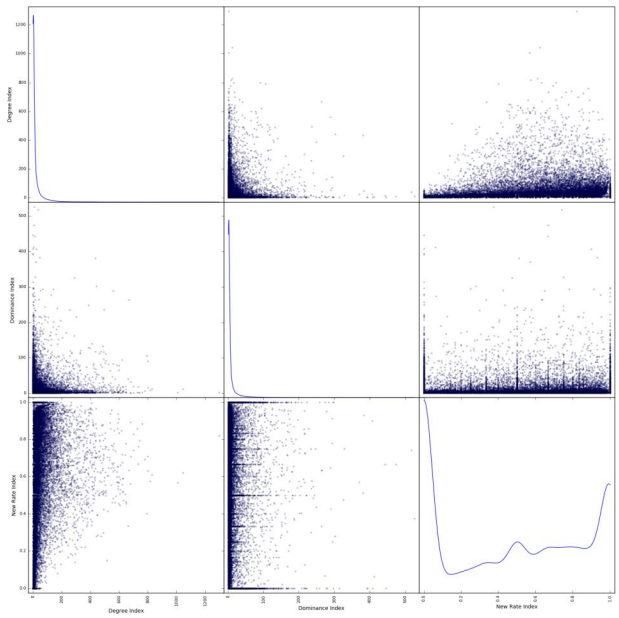


Figure 4: Scatter Matrix of Different Indexes for E_i

- We can also refer to [1] where several measures are described theoretically. For example, Shannon entropy might be an appropriate:

$$\text{Entropy Index} = -\sum_{U_j \in \text{nbr}_C(L_i)} p_{L_i}(U_j) \log p_{L_i}(U_j)$$

  where $p_{L_i}(U_j)$ is the probability that a given check-in at location L_i is made by user U_j. Shannon entropy measures the diversity/divergence of users visiting a certain venue. The more diverse visitors are, the more likely that they could meet each other here.

Similarly, we have the scatter matrix in Figure 5.

5.3 Relationship between E_i and B_i

We explore the relationship between these two indexes separately from the perspective of user and location. The main questions are: Do explorers tend to go to places that bring strangers together? Do places that bring strangers tend to attract explorers? Here we choose E_i as **Dominance Index** and B_i as **Entropy Index** to show the scatter plot in Figure 6. We also use

- For location L_i with B_i, we consider all the users that have checked in here and get the mean and median of their E_j scores. The Pearson Correlation and corresponding p-values are in Table 3, from which we can infer that there exists negative correlation. Therefore, one user that tends to explore are more likely to go to the places that bring non-strangers together.

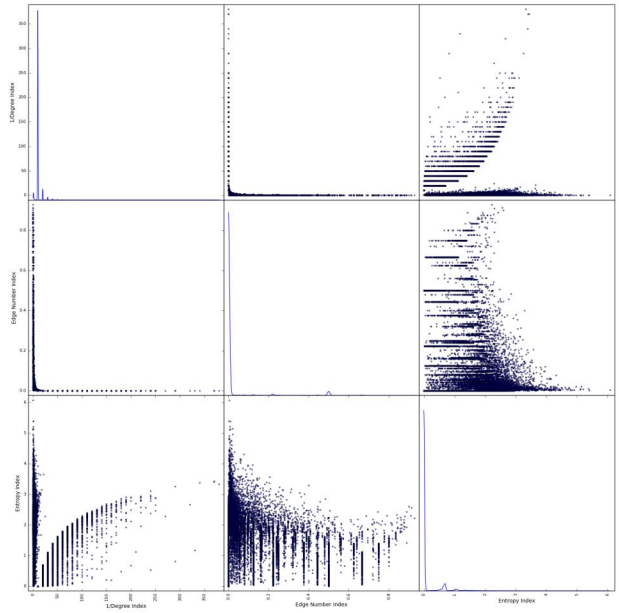
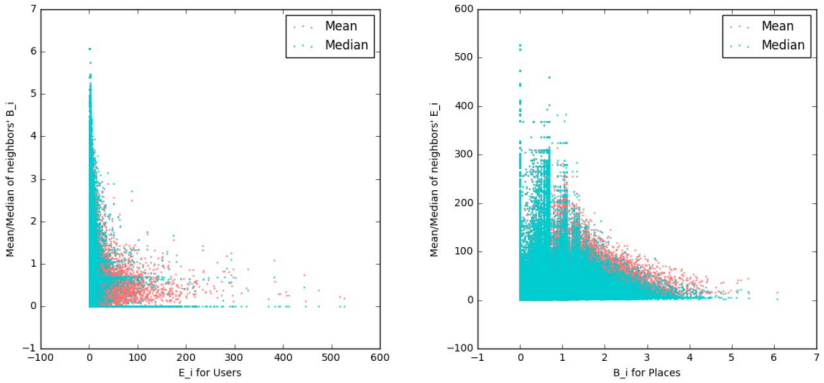


Figure 5: Scatter Matrix of Different Indexes for B_i



(a) Mean/Median($nbr_C(B_i)$) VS E_i **(b)** Mean/Median($nbr_C(E_i)$) VS B_i

Figure 6: Relationship between B_i and E_i

$E_i \setminus B_i$	Mean of Neighbors'			Median of Neighbors'		
	1/Degree Index	Edge Num Index	Entropy Index	1/Degree Index	Edge Num Index	Entropy Index
Degree Index	(-0.0052, 2.4e-3)	(-0.0213, 1.4e-6	(-0.0165, 1.8e-4)	(-0.0125, 4.4e-3)	(-0.0167, 1.6e-4)	(-0.0297, 1.5e-11)
New Rate Index	(0.0028, 5.3e-1)	(0.0073, 9.7e-2)	(-0.0102, 1.0e-2)	(0.0013, 7.7e-1)	(-0.0007, 8.6e-1)	(-0.0130, 3.19e-3)
Dominance Index	(0.0105, 1.7e-1)	(-0.0389, 1.2e-18)	(-0.1516, 5.1e-262)	(-0.0037, 3.9e-1)	(-0.0649, 3.6e-49)	(-0.1786, 0)

Table 3: (Correlation, p-value) for E_i and user's neighbors' B_i

$B_i \setminus E_i$	Mean of Neighbors'			Median of Neighbors'		
	Degree Index	New Rate Index	Dominance Index	Degree Index	New Rate Index	Dominance Index
1/Degree Index	(-0.0067, 2.9e-9)	(-0.0022, 5.8e-2)	(-0.0192, 8.1e-64)	(-0.0241, 2.1e-99)	(-0.0018, 1.2e-1)	(-0.0323, 4.3e-177)
Edge Num Index	(0.0375, 3.2e-238)	(0.0107, 4.2e-21)	(0.0407, 1.4e-280)	(0.0187, 1.2e-6)	(0.01246, 6.5e-28)	(0.0267, 3.2e-122)
Entropy Index	(0.0291, 9.9e-145)	(0.0067, 3.4e-9)	(0.0093, 3,2e-16)	(-0.027, 5.7e-128)	(0.0118, 3.1e-25)	(-0.0323, 2.6e-177)

Table 4: (Correlation, p-value) for B_i and location's neighbors' E_i

- For user U_i with E_i, we consider all the places he/she has checked in and get the mean and median of their B_j scores. The Pearson Correlation and corresponding p-values are in Table 4. There's no obvious positive or negative correlation, which means that venues with different B_i values seem to attract different types of people equally.

6 Summary

In this project, we specifically investigated the differet types of people's behaviors and venues' functions. We constructed three graphs including the graph of user friendship, the graph of geographical relations and the graph of check-in behavior.

Then we defined an index for users to measure the extent to which that they tends to explore more different places, and an index for venues to measure the extent to which that they tend to bring strangers together. We have multiple ways to express these indexes.

We also studied the relationship between users and venues. We found that an "explorer" is more likely to go to the places that bring non-strangers together. However, venues with different B_i values seem to attract different types of people equally.

We do have some limitations too. We didn't find a good way to evaluate if an index is good enough to represent the thing we want. As a compromise, we use them all when investigating the relations between human behavior and venue functions.

7 Individual Contributions

- **Chen Guo**: Problem formulation, coding up algorithms, plotting graphs, writing up reports.
- **Bowei Ma**: Data collection and cleaning, discussing about algorithms, making poster.

References

[1] Hristova, D., Williams, M. J., Musolesi, M., Panzarasa, P., & Mascolo, C. (2016, April). Measuring Urban Social Diversity Using Interconnected Geo-Social Networks. In Proceedings of the 25th International Conference on World Wide Web (pp. 21-30). International World Wide Web Conferences Steering Committee.

[2] Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., & Barabsi, A. L. (2015). Returners and explorers dichotomy in human mobility. Nature communications, 6.

[3] Cho, E., Myers, S. A., & Leskovec, J. (2011, August). Friendship and mobility: user movement in location-based social networks. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1082-1090). ACM.

[4] Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011). Socio-Spatial Properties of Online Location-Based Social Networks. ICWSM, 11, 329-336.

[5] Gao, H., Tang, J., & Liu, H. (2012, June). Exploring Social-Historical Ties on Location-Based Social Networks. In ICWSM.