L109 Assignment

Kaho Sato
King's College
ks789@cam.ac.uk

1 Introduction

2 Dataset

2.1 Foursquare Dataset

Foursquare dataset describes check-ins in New York City from XXX to XXX.The dataset consists of two tables. Each row of the first table describes a venue and contains the following information:

- id
- location represented as a Mercator coordinate
- category (e.g. 'Bakery', 'Apartment Building')
- total number of users who checked in
- total check-ins
- name (e.g. 'Le Pain Quotidien')

Each row in the second table describes a transition made by a user and contains the following information:

- id of the first venue
- id of the second venue
- time of the first check-in
- time of the second check-in

2.2 Airbnb Dataset

Airbnb dataset was compiled by Murray Cox from publically available information on Airbnb website to build *Inside Airbnb* (Online, b), a tool to explore how Airbnb is used in various cities, including New York City. The motivation of creating the tool was to demonstrate how Airbnb could be harmful to the residential housing. The dataset has been used to expose various other social issues such as gentrification (Gant, 2016) and digital discrimination (Edelman and Luca, 2014). There are four tables and each describes:

- listings
- when each listing is made available by the owner
- user reviews for listings
- name of the neighbourhoods

3 Graph

In this section, I specify a weighted undirected graph $\mathcal{G}$ constructed from Foursquare dataset and justify the design choices. Node in $\mathcal{G}$ represents a venue and an edge exists between two nodes if and only if the dataset contains a transition from one to the other. A possible interpretation of the edge is association between two places. The assumption is that, if one visits two places one after the other, there must be some relation between two places.

As stated previously, what I aim to extract from this graph is the notion of neighbourhood which consists of venues that people would "associate" with each other. This concept of association should be invariant of the order of the visit. Therefore I disregard the direction of the transition and $\mathcal{G}$ is an undirected graph.

The weight of the edge then should represent the strength of association between two venues. In this report, we assume that one associate two locations more strongly if they are geographically close to each other. Based on this assumption, one may define the weight of an edge to be an inversely proportionate to a positive increasing function of distance. Though the context is different, some works attempt to relate spatial distance with non-spacial concepts, such as friendship. Letting d to be the distance between two indivisuals, Backstrom et al. (2010) claims that the probability of them having a social connection is proportionate to d^{-1}. Others argue that it should be proportionate to d^{-2} (Lambiotte et al., 2008). Here we assume that the strength of the association between two places decays in proportionate to the inverse of the distance. For simplification, I define the distance between two venues to be the distance as the crow flies. One may elaborate this by, for instance, using how long it takes from one location to the other according to Google Maps.

It is also reasonable to assume that if two venues

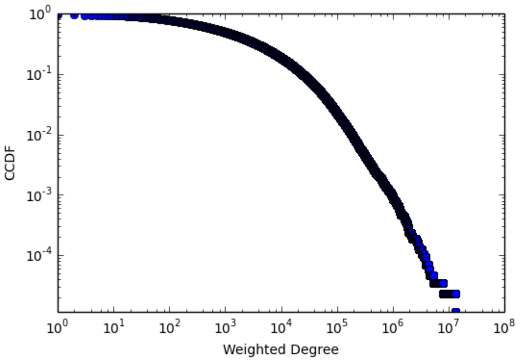

Figure 1: CCDF of weighted degree in $\mathcal{G}$

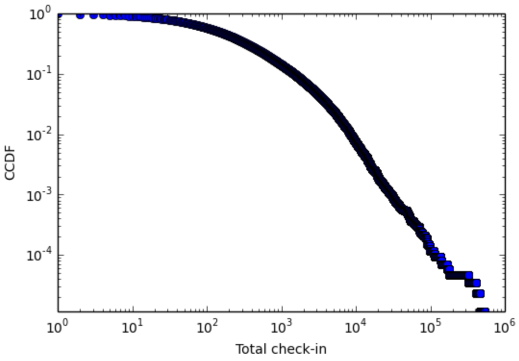

Figure 2: CCDF of total check-in counts in $\mathcal{G}$

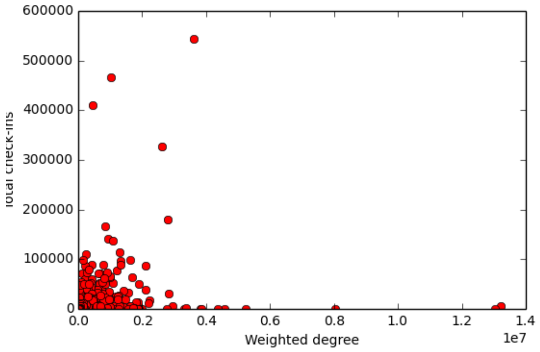

Figure 3: Relation between weighted degree and total check-in counts

are associated strongly, more transitions should be made between them. In this report we assume that the strength of the association between two venues is in proportionate to how many transitions are recorded.

Therefore I define the weight of the edge between node i, j to be:

$$w_{i,j} = \frac{transition\_count(i,j)}{distance(i,j)}$$

where $transition\_count(i,j)$ returns the number of edges between node i, j, and $distance(i,j)$ returns the distance between venues represented by node i, j.

4 Analysis

4.1 Degree distribution

Figure 1 shows the complementary cumulative distribution function (CCDF) of the proportion of nodes whose degree is larger than the given degree. As stated previously, the weight of an edge is an inverse of the distance between venues, which represent the strength of association between two venues. Then weighted degree is how strongly a venue is associated with all the other venues, therefore shows the relevance of a venue. The power law behaviour seen as the straight line in Figure 1 shows that the relevance of venues is largely heterogeneous and could be explained by rich-get-richer model. Indeed, it is perfectly plausible that the higher the relevance of a venue is, the more people visit the location.

One may say that the total number of check-in represents the popularity of a venue. Expectedly, this also exhibits the power law distribution as can be seen in Figure 2, which is CCDF of the proportion of nodes whose total check-in count is larger than the given count.

Relevance and popularity seems to be two concepts that could be related somehow. Unfortunately the data says otherwise. Each point in Figure 3 and Figure 4 represents a venue in Foursquare dataset, where the latter only shows the venue with fewer than 5000 check-ins. There does not seem to be any clear correlation between two metrics.

4.2 Weight distribution

Recall that we defined the weight of the edge to be an inverse of the distance between two nodes which it connects. Since distance is much more intuitive metric to consider, we look at the distribution of the inverse of the weight (i.e. distance) in this section. Figure 5 shows the CCDF of the proportion of nodes whose distance is larger than the given distance, and Figure 6 shows the frequency distribution of the distance. We can interpreted these graphs to show the relationship between the distance and the likeliness of checking in. Unlike

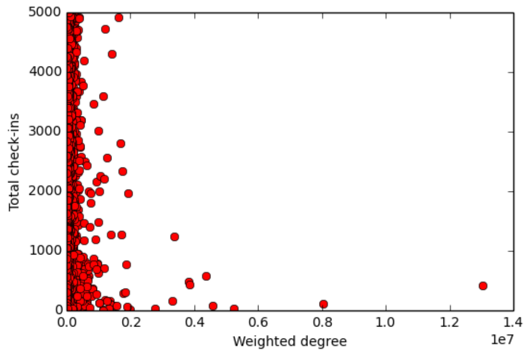

Figure 4: Relation between weighted degree and total check-in counts for venues with fewer than 5000 check-ins

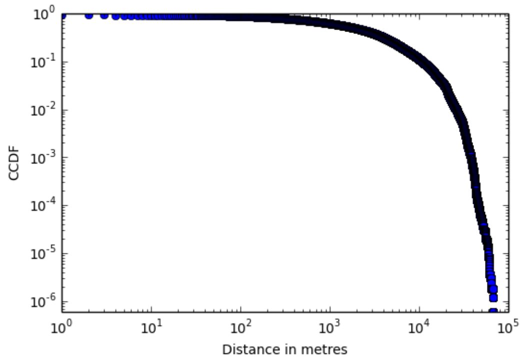

Figure 5: CCDF of the distance of the transition

the total check-in counts or the degree which follows the power law distribution, it shows an exponential distribution with a faster decaying tail. This is perhaps due to the fact that the check-ins in the dataset are limited to ones in New York, which imposes an upper limit.

4.3 Clustering Coefficient

A generalised clustering coefficient for weighted graph was prorposed by Saramäki et al. (2007). A weighted graph can be translated into a fully connected graph, where the edge that does not exist in the original graph bear zero weight. This gives the notion of *intensity* of a subgraph g with a set of edges e_g and a set of nodes n_g (Saramäki et al., 2007):

$$I(g) = \left(\prod_{(i,j) \in e_g} w_{i,j} \right)^{\frac{1}{|e_g|}}$$

An unweighted graph can be regarded as a weighted graph where all the edges bear a constant weight therefore the intensity of any triangle is constant. This can be generalised using intensity to obtain the clustering coefficient of a node i as follows(Saramäki et al., 2007):

$$C_i = \frac{2}{k_i(k_i - 1)} \sum_{j,k} (\widetilde{w}_{i,j} \widetilde{w}_{j,k} \widetilde{w}_{k,i})^{\frac{1}{3}}$$

where $\widetilde{w}_{i,j}$ is a weight of the edge between i and j normalised with respect to the largest weight of a edge in the network. The clustering coefficient of $\mathcal{G}$ is $7.99\text{e}{-5}$ which is very low. This is perhaps is because many triangles include at least one edge with very low weight, which pulls down the intensity of the triangle. The clustering coefficient of an unweighted graph $\mathcal{G}_u$, which has the same edges and nodes as $\mathcal{G}$ is 0.135. This is significantly higher than the clustering coefficient of an unweighted random graph $2.85\text{e}{-4}$with the same number of nodes and the average degree as $\mathcal{G}_u$. The clustering coefficient of $\mathcal{G}_u$ can be interpreted as the transitivity of the existence of association. In other words, it quantifies the extent to which venue A and C are visited in succession if venue A and B, and venue B and C are visited one after the other. average path length

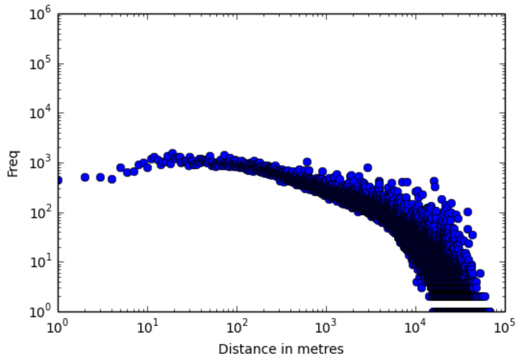

Figure 6: Distribution of the distance of the transition

5 Neighbourhood Analysis

Especially in a large city there is a certain notion of *neighbourhood*. The notion of neighbourhood is usually somewhat fuzzy, and often there is no well-defined geographical border between them. However, people do seem to have a general consensus on which venue belongs to which area. I argue that the notion of neighbourhood can be defined in terms of the association between venues. For instance, one neighbourhood where the majority of the population belongs to a certain socio-economic group would have many stores or restaurants which are well supported by such a group. We can say that these stores and restaurants should have a strong association among each other. Then, conversely, it must be possible to understand the general consensus of what the neighbourhood is from $\mathcal{G}$ by community detection assuming that the edge indeed models the association. We then see whether the definition of neighbourhood mined from $\mathcal{G}$ corresponds to neighbourhood defined by Airbnb.

5.1 Community Detection

In order to detect the community in $\mathcal{G}$, I used community detection module (Online, a) which implements the louvain method proposed by Lambiotte et al. (2008). This resulted to 4687 communities. Many consists of a very small number of venues, which are not big enough to be called a neighbourhood. In order to avoid the noise introduced by them, I filtered out the communities with fewer than 1000 venues. This resulted to 30 communities. Figure 7 shows Foursquare venues coloured according to the community they belong to. As can be seen, most of the venues which are geographically located closely to each other are in the same community, as the edge that connect them tend to have a higher weight. We can also observe some exceptions. These venue perhaps have at least one edge bearing a high weight with a venue that is geographically far, because there are multiple transitions between them. From here on, we refer to the communities detected as *Foursquare neighbourhood*, and neighbourhood defined by Airbnb as *Airbnb neighbourhood*.

5.2 Foursquare neighbourhood for Airbnb Listing

Let

$$S(l, n_{fsq}) = S(n_{fsq}) \cap k\_closest\_fsq(l, k)$$

where $S(n_{fsq})$ returns a set of Foursquare venues which belong to a Foursquare neighbourhood n_{fsq}, and $k\_closest\_fsq(l, k)$ returns k Foursquare venues which are closest to l. For a given Airbnb listing, I assign n_{fsq} such that:

$$assign\_fsq\_nbh(l) = \arg\max_{n_{fsq}} score(l, n_{fsq})$$

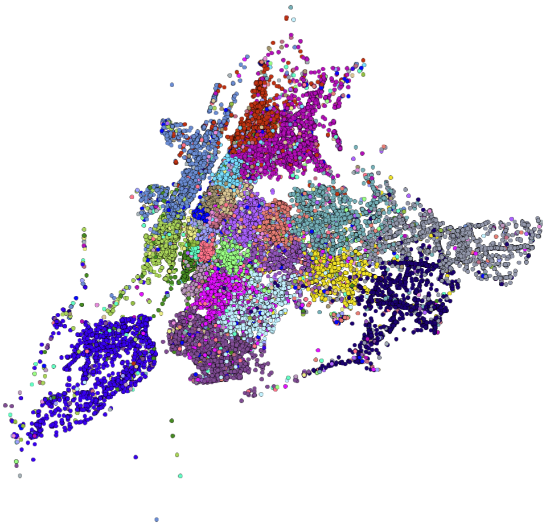

Figure 7: Foursquare venues coloured according to the community detected using the louvain method

where

$$score(l, n_{fsq}) = \sum_{v \in S(l, n_{fsq})} distance(l, v)^{-1}$$

Figure 8 shows the Airbnb listings coloured according to the Foursquare neighbourhood they were assigned to, and Figure 9 shows the listings coloured according to the Airbnb neighbourhood they belong to. The colour used for each Foursquare neighbourhood is the same for both Figure 7 and Figure 8. Note that Airbnb listings span over a much smaller region, which is why these figures look largely different from Figure 7.

5.3 Comparing Two Definitions

First observation one can make is that the Airbnb neighbourhood is more fine-grained with 230 neighbourhoods than the Foursquare neighbourhood with 30 neighbourhoods. This is not due to the threshold value of 1000 venues that I used to filter the neighbourhoods. With the threshold of 1 venue, there would have been 180 neighbourhoods which are already fewer than the number of the Airbnb neighbourhoods. This may though be related to the rate of the decay of the weight with respect to the distance. For instance, if I chose to define $w_{i,j}$ to be inversely proportionate to $distance(i, j)^2$, the Foursquare neighbourhood potentially would have been more fine-grained.

Recall that there are more Foursquare neighbourhoods. One can say that these two defini-

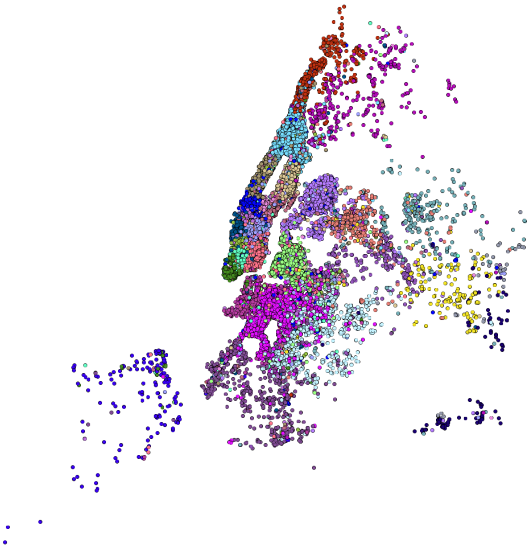

Figure 8: Airbnb listings coloured according to their Foursquare neighbourhood

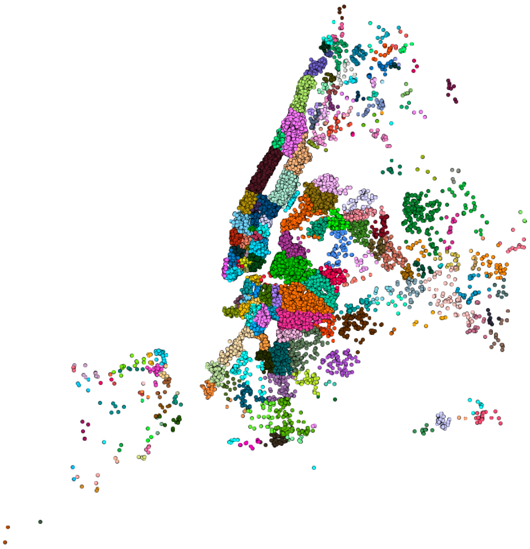

Figure 9: Airbnb listings coloured according to their Airbnb neighbourhood

tions have a high overlap if a given Foursquare neighbourhood is such that it groups up multiple entire Airbnb neighbourhoods without splitting them. To quantify the overlap, I first assigned on each Airbnb neighbourhood a Foursquare neighbourhood to which the highest number of its listings were assigned to:

$$assign\_fsq\_nbh(n_a) = \\ \arg\max_{n_{fsq}} |\{l | l \in listings(n_a), assign\_fsq\_nbh(l) = n_{fsq}\}|$$

where $listings(n_a)$ returns a set of Airbnb listings which belong to n_a. Then I computed $overlap$ for the set of all the Airbnb listings S_a:

$$overlap(S_a) = \\ \frac{|\{assign\_fsq\_nbh(l) = assign\_fsq\_nbh(nbh\_a(l)) | l \in S_a\}|}{|S_a|}$$

where $nbh\_a(l)$ returns the Airbnb neighbourhood that l belongs to.

As a result, I found that $overlap(S_a) = 0.653$, which is considerably higher than $3.33\text{e}{-02}$ which what it would have been if a Foursquare neighbourhood was assigned at random on a listing. From this, we can conclude that Foursquare neighbourhood is such that it groups multiple Airbnb neighbourhoods and two definitions of neighbourhood somewhat overlaps.

5.4 Popularity

The total number of check-in is an indicator of the popularity of a Foursquare venue (Noulas et al., 2011). One may consider an area to be a large Foursquare venue and naturally quantify its popularity as a sum of the number of check-in for Foursquare venues which belong to it. Let this measure of popularity of an area n be $p_{fsq}(n)$. On the contrary, there is no agreed way to quantify the popularity of an area using the information about Airbnb listings. In this section, I investigate various statistics of the Airbnb listings which belongs to an area, and see if any of them could be as good as $p_{fsq}(n)$ to quantify the popularity. I do this by computing the statistics on a Foursquare neighbourhood n and measuring the correlation with $p_{fsq}(n)$.

5.4.1 Candidate Statistics

Intuitively, a listing is popular if there is more activity; that is, if it is occupied more frequently. Unfortunately, the Airbnb dataset does not offer such information. In the previous works, the activity of a listing was estimated by the number of reviews per month (Cansoy and Schor, 2016; Online, b). Airbnb guests may leave a review after their stay. Although it is optional and therefore not all the guests do so, this may be used as an indicator of Airbnb activity.

Another possible indicator of the popularity of a listing would be the monthly income it generates. Again, this information is not available from the dataset. In the previous works, the minimum income per month was used instead. This can be computed as the product of the minimum length of stay, price and the reviews per month (Cansoy and Schor, 2016; Online, b).

I propose the following statistics to measure the popularity of an area:

1. $p_{a1}(n)$
   Average number of reviews per month given to the listings in an area n
2. $p_{a2}(n)$
   Total number of reviews per month given to the listings in an area n
3. $p_{a3}(n)$
   Average minimum income per month among the listings in an area n
4. $p_{a4}(n)$
   Total minimum income per month earned by all the listings in an area n

5.4.2 Results

Figure 10, 11, 12 and 13 are scatter plots where each point represents a Foursquare neighbourhood. The x-coordinate represents its total check-in counts for the venues and the y-coordinate represents the respective candidate metric proposed in 5.4.1.

There is no clear correlation between $p_{a1}(n)$ that can be read from Figure 10. The other statistics, though heteroskedastic, suggest a linear correlation with $p_{fsq}(n)$.

Table 1 shows Pearson's correlation coefficient r and p-value p between each candidate proposed in 5.4.1 and p_{fsq}. Taking the threshold value of 0.01, only p_{a3} and p_{a4} are significantly correlated to p_{fsq}. From this we can conclude that a good statistic which can be obtained from the Airbnb dataset to measure the popularity of an area is either p_{a3} or p_{a4}, though the lower p-value for p_{a3} suggests that p_{a3} is a better statistic for this purpose, assuming that p_{fsq} is indeed a good measure of the popularity of an area.

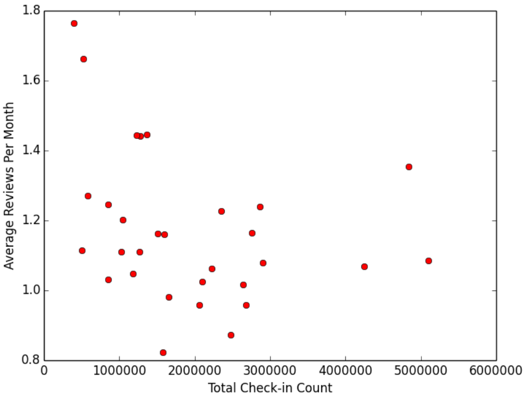

Figure 10: Total Check-in Count and Average Reviews Per Month of Foursquare Neighbourhoods ($p_{a1}(n)$ against $p_{fsq}(n)$)

There are two possible reason why p_{a3} and p_{a4} are better suited for measuring the popularity of an area than p_{a1} and p_{a2}. One is that the number of reviews per month of a listing may be influenced much more than the popularity of the area, such as the demographics or the purpose of the visits of the guests.

Though the minimum income per month is not independent of these factors, as it is after all defined in terms of the number of reviews per month, the noise introduced might be offset by other variables introduced, which are the price and the minimum length of stay, which possibly has a high correlation with p_{fsq}.

Figure 14 and 15 are scatter plots where each point represents a Foursquare neighbourhood. The x-axis is again the total check-in counts of the venues in the neighbourhood and the y-axis is the average price of the and the average minimum length of stay of a listing, respectively. Unsurprisingly, the average minimum length of stay seems to be invariant of p_{fsq}. The average price though seems to have a linear correlation with p_{fsq}. Indeed Pearson's correlation coefficient between the average price and p_{fsq} is 0.742 with p-value of 2.68e$-$06, which is lower than that of p_{a4}. From this we can conclude that the average price of the listing is the best indicator of the popularity of the area. This result goes well with the intuition that a guest would be willing to pay more to be in a popular area.

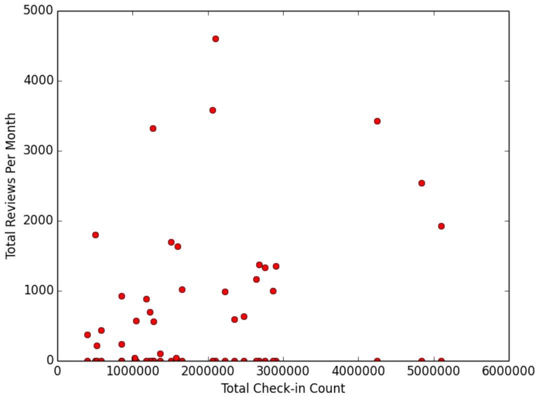

Figure 11: Total Check-in Count and Total Reviews Per Month of Foursquare Neighbourhoods ($p_{a2}(n)$ against $p_{fsq}(n)$)

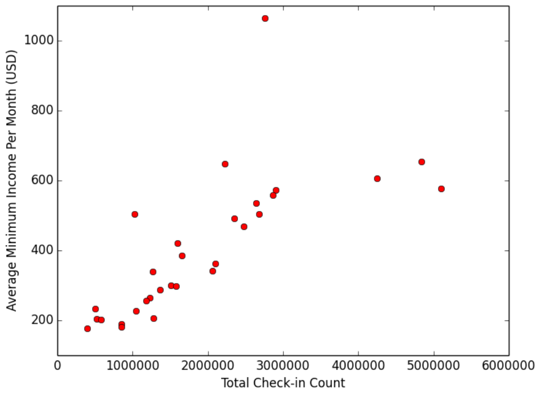

Figure 12: Total Check-in Count and Average Minimum Income Per Month of Foursquare Neighbourhoods ($p_{a3}(n)$ against $p_{fsq}(n)$))

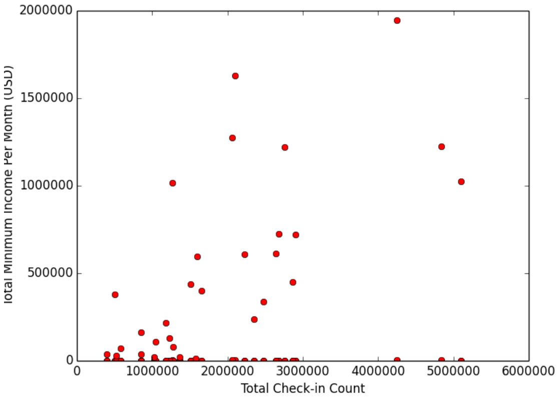

Figure 13: Total Check-in Count and Total Minimum Income Per Month of Foursquare Neighbourhoods ($p_{a4}(n)$ against $p_{fsq}(n)$)

Table 1: Pearson's Correlation Coefficient r and p-value p

	r	p
p_{a2}	0.439	1.53e−02
p_{a3}	0.728	5.07e−06
p_{a4}	0.697	1.88e−05

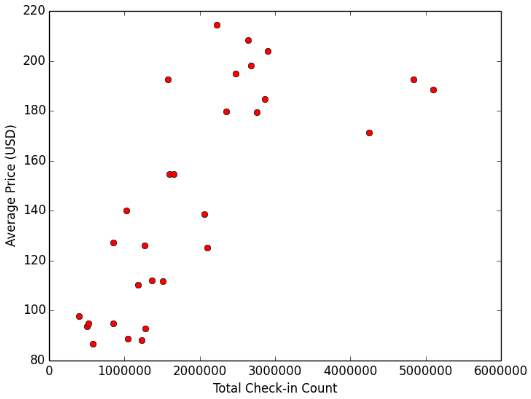

Figure 14: Total Check-in Count and Average Price of Listings in Foursquare Neighbourhoods

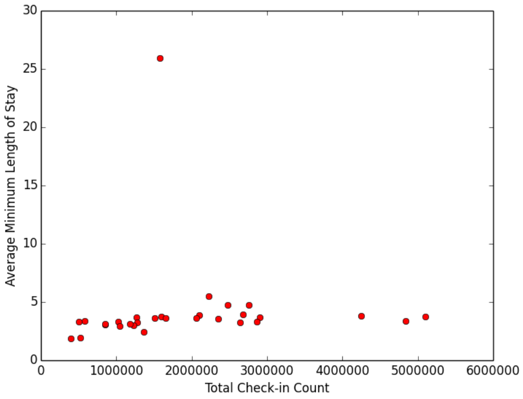

Figure 15: Total Check-in Count and Average Minimum of Length of Stay of Listings in Foursquare Neighbourhoods

6 Discussion

References

Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

Mehmet Cansoy and Juliet Schor. Who gets to share in the sharing economy: Understanding the patterns of participation and exchange in airbnb. *Unpublished Paper, Boston College*, 2016.

Benjamin G Edelman and Michael Luca. Digital discrimination: The case of airbnb. com. 2014.

Agustín Cócola Gant. Holiday rentals: The new gentrification battlefront. *Sociological Research Online*, 21(3):10, 2016.

Renaud Lambiotte, Vincent D Blondel, Cristobald De Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387 (21):5317–5325, 2008.

Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. *ICwSM*, 11:70–573, 2011.

Online. Community detection for networkx. `http://perso.crans.org/aynaud/communities`, a. Accessed:2017-03-03.

Online. Inside airbnb. `http://insideairbnb.com`, b. Accessed:2017-03-03.

Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertesz. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):027105, 2007.