1. Results of proposed method

New York	Accuracy	precision	recall	F1-score
Method 1	99.8%	19.2%	6.9%	10.1%
Method 2	99.9%	11.4%	6.5%	8.3%
Method 3	99.8%	8%	9.6%	8.7%
Method 4	99.9%	19.2%	13.7%	16%
Method 5	99.8%	11.3%	9.2%	10.1%
Combined method	99.9%	16.6%	7.8%	10.6%

Paris	Accuracy	precision	recall	F1-score
Method 1	99.5%	56.4%	9.1%	15.7%
Method 2	99.6%	48.2%	9.5%	15.9%
Method 3	99.2%	9.9%	9.7%	9.8%
Method 4	99.6%	56.2%	18.2%	27.5%
Method 5	99.5%	34.8%	10.9%	16.6%
Combined method	99.5%	34.7%	11.1%	16.8%

2. What are the physical meanings of your proposed methods? Why do you want to do that?

   Method1

   利用每個人的打卡習慣(次數)及同地點前後半小時內同時打卡的次數計算兩兩之間的相似度，將相似度大於平均值的 pair 視為朋友。

   Method2

   假設平日白天打卡的位置為工作地點，同時同一工作地點打卡的人可能是同事關係，同事關係即有滿高的機率同時為朋友關係。因此，利用每個人的打卡習慣(次數)、同地點於平日白天前後一小時內同時打卡的次數計算兩兩之間的相似度，將相似度大於平均值的 pair 視為朋友。

   Method3

   嘗試不仰賴時間因素，只考慮打卡地點之間的距離及打卡地點的熱門程度計算兩兩之間的相似度，引用 paper “Distance and Friendship: A Distance-based Model for Link Prediction in Social Networks” published by Yang Zhang and Jun Pang，利用距離及打卡地點的熱門程度來區分朋友關係與陌生人關係。

   計算方法如下:

$$locent(\ell) = -\sum \frac{|ci(\ell,u)|}{|ci(\ell)|} \log \frac{|ci(\ell,u)|}{|ci(\ell)|}$$

$locdiv(\ell) = \exp(locent(\ell))$ 此式表示地點的熱門程度

$$ldpd(u, u') = \{d(\ell, \ell') \cdot \max(locdiv(\ell), locdiv(\ell')) \mid \forall(\ell, \ell') \in m(u) \times m(u')\}.$$

Paper link: http://satoss.uni.lu/members/jun/papers/APWeb15.pdf
兩地距離與地點的熱門程度相乘的結果愈小愈有機會是朋友關係，然而用這個方式針對每對 pair 一一計算執行時間會過長，所以我取了 Method1、Method2、Method4、Method5，以及 Failed Method 來當基準 pair，去計算這些 pair 的 ldpd，取較小的 pair。

Method4

假設兩人雖無同時打卡，但擁有共同朋友，就有很大的機率也是朋友，因此利用 Method1 得到的朋友關係結果，將有共同朋友的兩人都視為朋友。

Method5

假設兩人在熱門景點同時打卡是朋友的機率小於在非熱門景點同時打卡是朋友的機率。引用 paper “Distance and Friendship: A Distance-based Model for Link Prediction in Social Networks” published by Yang Zhang and Jun Pang，利用此篇中地點熱門度的計算方式，算出前 25 大熱門景點，將同天同地點在非 25 大熱門景點的兩人視為朋友。

Combined Method

利用前面 5 個 method 的結果，進行是否為朋友的投票，高於 2 票則為朋友。

Failed Method

計算每個人打卡地點的平均經緯度，再算出兩兩之間的的距離，愈小的愈可能是朋友，結果滿差的。

3. Do you feel that these methods will have good performance in San Francisco, why?
   San Francisco 的打卡資料較為龐大，可能隱藏的變數也更多。若考慮打卡時間的相似度，我認為仍會有一定的準確度，但若要以距離及打卡地點的熱門度來預測，performance 應該會下降許多。
4. If same approach results in different performance among cities, what’s the possible reason?
   打卡資訊隱藏了許多因素，都市人與郊區的居民打卡習慣就會有所不同，以熱門景點為例，每個城市的熱門景點集中的程度就有差異，一座城市觀光客的比例也會讓打卡資訊有所不同。從打卡資料中，我還觀察到幾秒內，同一個人就有可能在不同地點打卡，這很有可能是人們在造訪一地之後回到家裡做補打卡的動作，這樣的比例一高，我們也很難從打卡資料中預測出人們的移動路徑。每個人的打卡習慣不同，不同城市的居民又會有更大的不同呀!