# Examination of Business Metrics in Online Geosocial Networks

Amit Chattopadhyay (amitch@stanford.edu)

## INTRODUCTION

Over the last few years, there has been a tremendous increase in information sharing about personal activities that makes people the authoritative voice of information and giving power to their actions that has never been seen previously in the history of humankind. The unprecedented growth of online social networks is primarily due to the power that it gives users when sharing their life with their friends. While there is no doubt that connectivity of people within these social networks has improved considerably, it has also allowed for increased interest in bringing greater accessibility to niche culture, word of mouth propagation, viral growth of amateur music and videos that would never have made its way into mainstream media if it had not been for the power of online social networks. This increased power of social networks in propagating and making things viral is of particular interest to advertisers and merchants as they need to bring a fresh outlook to examine the new world media.

Of particular interest are online location based social networks like Foursquare and Facebook Places. They are a fascinating twist to social network information sharing because they have direct relevance to not just online advertising but also to physical shops. Users share their mobility information in online location based sharing networks as part of a procedure called 'check-in' and this provides a wealth of temporal data about the locations of interest of people in the network.

## MOTIVATION AND PROBLEM DEFINITION

This temporal data is of great interest to advertisers and retailers as the data can be harvested for many interesting pieces of information which can help guide business decisions in the rapidly devolving economy where online retailers have a clear advantage over physical shops. It can also be used by startups of geosocial networks to strengthen their case for improved relations with retail shops if it can be demonstrated that there is significant impact from such networks.

First we create a framework to perform analysis by creating a dataset of a geosocial network that can be analyzed inside a graph analysis tool like Snap. We examine the friendship network of this collected data set to ensure that we are looking at a social network that follows the norms of power laws. Then we calculate two important metrics of this study:

- The average extent of influence a location has in the social network graph
- The average extent of cascade that people in a geo-social network provide when checking into a location.

These questions lead to interesting discussions on how viral marketing can take place in such networks and we illustrate the business significance of these metrics that can be used to creatively market to an online audience.

In this study we have sampled data in the *foursquare* network to develop the social graph and the temporal data for forming the framework for answering these questions.

## BACKGROUND AND RELATED WORK

Analyzing the mobility patterns has long attracted attention of experts from ubiquitous computing, spatial data mining and statistic fields. In an analysis of 100,000 cellphone users trajectories ("Understanding individual human patterns") demonstrated that human mobility produces simple reproducible patterns. Related work has been taken up to model spatio-temporal patterns present in human mobility that can be used to create a model for predicting future location. Unlike cell phone data, checkins are unique in their social aspect. They

provide more information due to the nature of the friendship network they exist in as well as the specifics of the intended location of visit.

In "Friendship and mobility: user movement in location-based social networks", Leskovec et all took the data from a geosocial network and extended the trajectories study to analyze if a model can be generated for predicting locations in an online geosocial network. The study made some remarkable finds in the area of geographical movement, temporal dynamics and the social network which are used to help form a framework in our study. This study utilizes some of the same techniques described to ascertain the 'work' and 'home' location of a user in a network given the dataset of checkins. Of particular interest that is related to this study is the dynamics of the checkin distances. In this study it was found that people rarely checkin beyond 100kms from their home. In our study we delve into deeper depth examining what is the average range of checkins that a location receives. This is of particular interest to identify what are the dynamics of location in comparison to checkins. Most other works examine the opposite relation.

Specific works on examining Foursquare dataset have been done by Noulas et all in "An empirical study of geographic user activity patterns in foursquare". Although the transition activity patterns are not interesting for this study, the fact that the data collection employed for getting the Foursquare data set was of particular interest.

As part of investigating the question of influence cascades of checkins in the Foursquare network, the work in "The dynamics of viral marketing" offered a lot of ground work for doing this study. In particular that study examined an Amazon network and the viral cascade of recommendations. In our study, we perform the same study in the Foursquare network, substituting 'recommendations' for 'checkins' that are visible to immediate friends. So a checkin visible to an immediate friend is like a 'recommendation'.

## METHODOLOGY

To perform the study, a collection of checkins were required. Initially the idea was to get the dataset from existing studies due to the limited time available for gathering the dataset. However due to many privacy issues the datasets could not be obtained. For the purpose of testing and study, we created our own dataset (with the original intention being to replace it with a much larger dataset once it was available). Since personal checkin information on Foursquare network is restricted to immediate friends, an indirect sampling approach was employed. Foursquare checkins allow users to publicly share the checkin over twitter as a status message. These status messages have a unique status text (which include the text – "4sq.com") which can be used to locate checkin data. The Twitter Search APIs are rate limited, so we employed the Twitter Streaming APIs to trawl through the status updates. For each user id that was collected, we also used the Foursquare APIs to construct a social graph of the Foursquare network. The feed crawler was employed for a total of 21 days generating 9,471,336 checkins for a total of 102,032 users all over the world. The tweets were then converted to a format that can be processed easily with the following fields (Time, UserId, VenueId, LocationCoordinates). In addition there were tables to cross reference the UserId (and related friends) which was collected by using the Foursquare APIs to search all friends of given UserId. In addition for each VenueId cross reference table we stored the GPS coordinates.

**Checkin Distribution**

**(21 days study)**

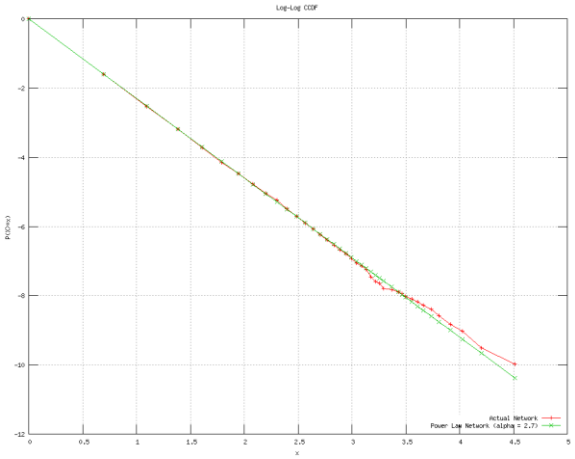
Total Checkins: 9,471,336
Total Users: 102,032

Checkins in US: 6,321,456
Total Users in US: 45,063

(Seattle Area)
Filtered sample of Checkins: 151,543
Filtered sample of users: 2,564

First we created a graph for the friendship network in the Foursquare network sampled – each edge represents a friend relation listed from the FourSquare API and each node represents a user sampled.

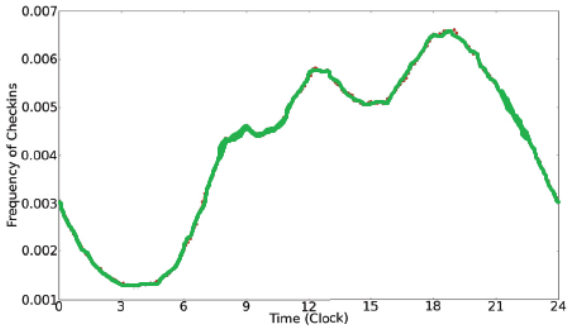
The degree distribution of the Foursquare friendship network of a sample 100,000 users in the networks (includes users from checkins and their friends mined using the API) was found to be heavy tailed with majority of them being low degree nodes and only few with significantly larger degrees confirming that this is a power law network.



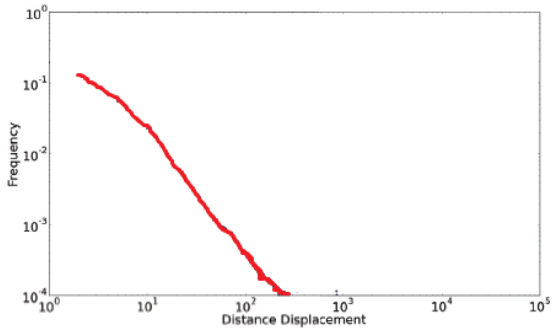
Next, we filtered the dataset to only include users who were only within the Seattle area. This was required for the framework as we thought including the international dataset and of other states would simply make the scope of the project too big to complete in the limited time frame. Also each region may need to be studied separately to be able to explain many macro phenomenon that are specific to a region. In the end the data was reduced to a set of 151,543 checkins and 2,564 users who had tweeted their checkin status in the Seattle area.

For each user we also went about figuring out algorithmically their 'home' location by averaging the data points of all their checkins. This was required for calculating for the later part of the study which requires us to measure distance between the user and the checkin.

The data set revealed that most checkins happened in the evening, presumably from night life.

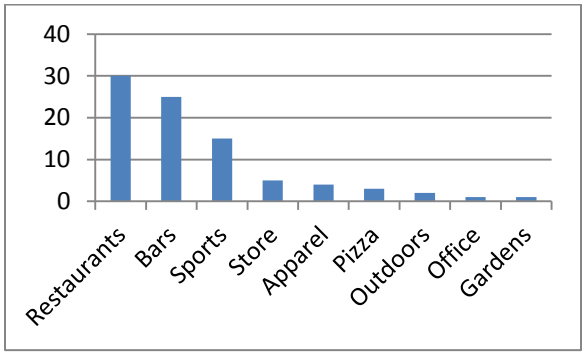



Given that we had calculated the 'home' location by averaging the data points, we found the average distance displacement of users and plotted as follows. This shows roughly that checkins beyond 100 miles of a 'home' location was pretty rare with the average being around 1-5 miles of the 'home' location.




We also plotted the frequency of checkins to different types of establishments which gives a good idea that our study which is aimed at primarily targeting business metrics caters to the retail market

(and not checkins to offices/outdoors, which would then be not useful for our study).



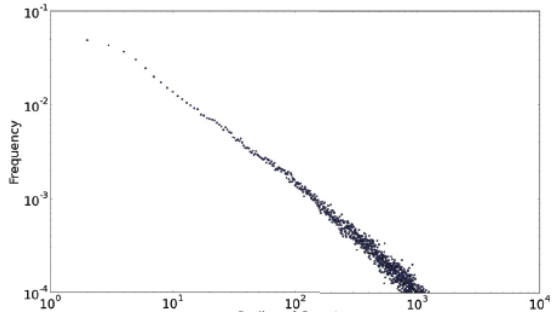

Now that we had a framework of data and validated some key aspects of the data set – primarily power law of the friendship network and the fact that checkins that had most data points were retail related, we were ready to answer the first of our two questions – the average distance of influence a business establishment can exhibit.

This measure of radius of influence of a location is:

$$\text{Radius of Influence} = \sqrt{\frac{1}{n}\sum_1^n (r_i - r_u)^2}$$

which is the standard deviation of distance of a location and every user's 'home' location when checking into the location. Here r(i) is the location of the establishment, r(u) the home location of the user that checked into the location i. A low radius of influence would indicate that the checkin location has a very 'local' or 'niche' influence.

Plotting the binned radius of influence on a log-log scale gave us the distribution graph of the radius of influence.




which follows a power law.

Of further interest would be to break down this graph and create it for each location type and present the finding. Due to time constraints this exercise was not performed on this data set.

Next, we answer the question of influence cascade exhibited by a user when checking into a location, for which the temporal data had to be analyzed. But first we describe the model used to describe the influence in foursquare checkins.

**Diffusion Model of Foursquare Checkins**

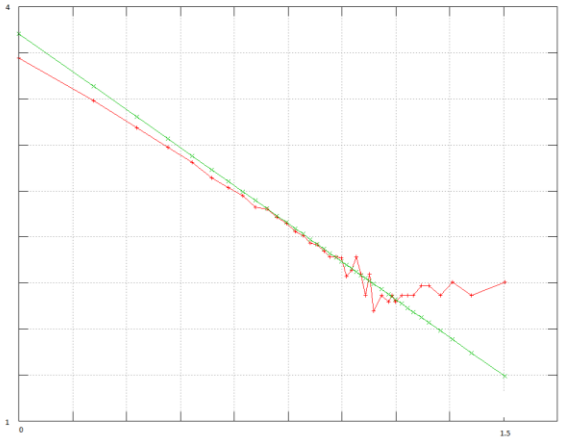
In our model we use the *Cascade model* which says that whenever a neighbor v of node u adopts, then node u also adopts with probability p(u,v). In other words, every time a neighbor of u checks in to a location, there is a chance that u will also check in at some later point.

We represent the temporal data of each checkin as a directed graph. The nodes represent users and a directed edge indicates information about the checkin that was passed on to the neighbor. The tuple (I,J,P,L) represents users I and J where I had a checkin to location L which showed up in J's friend's history and was then at some later point of time executed by checking into L.

The process of generating edges is as follows, a node I first checks into location L at time T and has this checkin show up in all his friends j(1)…j(n). Then j(i) checks into the same location and this cycle repeats. If multiple friends of j(i) show up checking into location L, the first neighbor to checkin to L is given priority.

In order to identify cascades, the causal propagation of checkin at a location, we track checkins that were successfully adopted. A checkin is successfully adopted if the user sees a neighbors checkin and later checks into the same location continuing the cascade. This is only for the first time that person checked into that location.

A plot of the cascade is presented below.



## RESULTS

This study examined a large dataset of a geosocial network and examined two metrics. We found the Radius of Influence of checkins follows the power law and that 70% of checkins were in the 1-5 mile radius of the checkin locations radius. The radius of influence also follows a power law with a heavy tail. It would be also interesting to drill into this data and classify by checkin location type.

Further, the examination of the cascade of checkins in the social network we noticed that most cascades have limited length and quite frequently do not exceed more than 3 to 4 hops. This indicates that viral influences may not be very effective. After examining this result it was concluded that the data set might need to be over a much larger time period and with more information about a specific region to conclusively conclude this finding. The problem with the current data set is that being limited to 3 weeks of data, the cascades may not be showing up correctly.

A few interesting things to do with this data is to compute the radius of influence hotspots and overlay on a map to get a better picture of relative competitiveness of checkin locations of the same type. This can be used to assess if a certain business is doing better than a similar counterpart. Also, the cascade information for specific business types can be used to understand better correlation between what type of establishments are more likely to be visited due a friend's checkin through the social network.

In conclusion, this study laid the framework for two business metrics that can be used as a basis for assessing the relative strength of range of influence. In addition the examination of the cascade range can be used for future studies on predicting viral influences in geosocial networks.

## REFERENCES


*Leskovec, J. and Adamic, L.A. and Huberman, B.A.* 2007. The dynamics of viral marketing. ACM

*Centola, D*. 2010. The spread of behavior in an online social network experiment. American Association for the Advancement of Science

*Cho, E. and Myers, S.A. and Leskovec, J*. 2011. Friendship and mobility: user movement in location-based social networks. ACM

*Noulas, A. and Scellato, S. and Mascolo, C. and Pontil, M*. 2011. An empirical study of geographic user activity patterns in foursquare. ICWSM'11

*Goyal, A and Bonchi, F. and Lakshmanan, V.S.* Learning Influence Probabilities In Social Networks.

*S. Scellato, C. Mascolo, M. Musolesi, and V. Latora.* Distance Matters: Geo-social Metrics for Online Social Networks. In Proceedings of the 3rd Workshop on On-line Social Networks (WOSN 2010), June 2010.

Gonzalez, M.C. and Hidalgo, C.A. and Barabasi, A.L. 2008. Understanding individual human mobility patterns. Nature Publishing Group

*Zheng, Y. and Zhang, L. and Xie, X. and Ma, W.Y*. 2009. Mining interesting locations and travel sequences from GPS trajectories. ACM

*Kleinberg, J, 2000.* The small-world phenomenon: an algorithm perspective, Proceedings of the thirty-second annual ACM symposium on Theory of computing.

*Clauset, A. and Shalizi, C.R. and Newman, M.E.J. 2007.* Power-law distributions in empirical data. Arxiv preprint arxiv:0706.1062

*Eubank, S. and Guclu, H. and Kumar, V.S.A. and Marathe, M.V. and Srinivasan, A. and Toroczkai, Z. and Wang, N.* 2004. Modelling disease outbreaks in realistic urban social networks. Nature Publishing Group

*Watts, D.J.* 2002. A simple model of global cascades on random networks. National Acad Sciences