Location-Based Influence Maximization in Social Networks

Tao Zhou, Jiuxin Cao*, Bo Liu, Shuai Xu, Ziqing Zhu, Junzhou Luo
Key Laboratory of Computer Network and Information Integration of MoE of China(No.93k-9)
School of Computer Science and Engineering
Southeast University, Nanjing, China
{zhoutao, jx.cao, bliu, xushuai7, zzqxztc, jluo}@seu.edu.cn

ABSTRACT

In this paper, we aim at the product promotion in O2O model and carry out the research of location-based influence maximization on the platform of LBSN. As offline consuming behavior exists under the O2O environment, the traditional online influence diffusion model could not describe the product acceptance accurately. Moreover, the existing researches of influence maximization tend to only concern on the online network of relationships but rarely take the offline part into consideration. This paper introduces the location property into the influence maximization to accord with the characteristic of O2O model. Firstly, we propose an improved influence diffusion model called TP Model which could accurately describe the process of accepting products under the O2O environment. Meanwhile, the definition of location-based influence maximization is presented. Then the user mobility pattern is analyzed and the calculation method of offline probability is designed. Considering the influence ability, a location-based influence maximization algorithm named TPH is proposed. Experiments prove TPH algorithm has general advantage. Finally, focusing on the performance of TPH algorithm under special circumstances, MR algorithm is designed as complement and experiments also verify its high effectiveness.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Sociology; H.2.8 [**Database Management**]: Database Application—*Data Mining*

General Terms

Algorithms, Experimentation, Performance

Keywords

Diffusion model, influence maximization, LBSN, O2O, social networks

*Contact Author

CIKM'15, October 19–23, 2015, Melbourne, VIC, Australia.

DOI: http://dx.doi.org/10.1145/2806416.2806462.

1. INTRODUCTION

In recent years, social networks become increasingly popular providing important platform for people to share their ideas, activities and interests. Messages and ideas propagate fast through social networks, making it an outstanding environment for the promotion of products and services. Influence maximization is one of the major problems in social networks. It comes from the viral marketing which takes advantage of "word-of-mouth" effect to promote products. Influence maximization could be described as an algorithm problem of determining a certain number of initial users in social network to maximize the spread of influence through the word-of-mouth effect [1].

Existing researches on influence maximization have made significant achievement. However, they tend to focus only on the online social network, against the trend that the business model becomes more likely to connect the online and offline aspects together. The concept of a new business model called O2O draws increasingly more attention of the whole industry. O2O stands for Online to Offline, which means to conduct online promotion and purchasing in order to drive the offline marketing and consuming. The key point is to bring the online consumers to the offline shops. Rather than the physical goods in the traditional B2C model, the transaction subject in O2O model is the life service product such as catering and entertainment.

O2O model brings new challenges to the influence maximization. As we mentioned above, the traditional influence maximization problem lacks the concern about the offline parts of the product promotion and the traditional online influence diffusion model could not offer accurate description about the process of product acceptance any more. The factors that affect the diffusion process are not only the online features like the network topology, but also the offline user properties, because the daily activity area and consuming location preference will greatly affect the probability of accepting the life service product which has location property as the same. Meanwhile, the target of the influence maximization is altered, since the promotion needs to pay attention to the users who are more likely to carry out the consuming behavior in the aspect of the geography for gaining higher benefit. To sum up, the offline phase of the marketing should be regarded as important content when it comes to the O2O environment and we need to consider the influence diffusion model and the influence maximization problem in a location-based way.

O2O model has a close relation with the location-based social network (LBSN). Being different from the traditional

one, LBSN could track and share user location information in addition to providing communication approach for the user. There is a new kind of social structure generated by the user location in LBSN [2]. Carrying plenty of information of both social network and location-based behavior, LBSN becomes a nonnegligible platform for the product promotion in O2O model. So for the location-based influence maximization, LBSN would be an important research environment.

In this paper, we will focus on the influence maximization in O2O model based on the real data of LBSN. A new diffusion model for O2O will be designed combining the online and offline parts and the definition of location-based influence maximization problem will be given. We will consider the user location property and make use of historical location information to explore the location preference of user offline consuming behavior. Finally two algorithms called TPH and MR will be presented and the experiments prove the algorithms effective in solving the location-based influence maximization problem.

The remainder of the paper is organized as follows. In Section 2, the related work is presented. The dataset we use in this paper is introduced in Section 3. A new influence diffusion model is proposed and the location-based influence maximization problem based on the model is defined in Section 4. In Section 5, online probability, an important parameter in the model and the problem, is discussed and the calculation method is introduced. Two algorithms for the problem are designed in Section 6 and the performance of them is evaluated. Finally, Section 7 concludes the paper.

2. RELATED WORKS

Influence maximization was first described as an algorithm problem by Domingos and Richardson [1], and was first formulated as a discrete optimization problem by Kempe, Kleinberg and Tardos [3]. Kempe et al. proved that it is an NP-hard problem and gave a greedy optimization algorithm with provable approximation guarantee. However, the greedy algorithm executes monte carlo simulation [4] to obtain the approximated solution of the influenced set size, which faces the shortage of high time complexity.

Some researches aim to improve the greedy algorithm. Leskovec et al. proposed the CELF (Cost-effective Lazy Forward Selection) algorithm [5]. It takes the advantage of submodularity property of influence maximization and greatly reduces the calculation of approximation. Goyal et al. [6] presented CELF++ algorithm which is an improvement of CELF, further decreasing the amount of calculation. Chen et al. [7] proposed two improvements of greedy algorithm: NewGreedy and MixedGreedy algorithms. The former one removes the edges which have no contribution to the influence diffusion to get a smaller graph and the latter one combines CELF with NewGreedy.

Another direction of research is the heuristic algorithm. Compared with greedy algorithm, it has the huge advantage in speed. Basic heuristic algorithms include Max Degree algorithm (selecting nodes according to the degree), distance centrality algorithm (selecting nodes according to the average distance to the other nodes) and so on. In [7], Chen et al. designed the Degree Discount Heuristic algorithm which makes discount to the degree of the neighbors when a node is selected into initial set. Beside the heuristic algorithm using network structure, there are other methods introducing the web search approach into node selecting, such as PageRank [8] and HITS [9] algorithms.

The influence maximization related with location is still a domain remaining to be researched. Li et al. [10] worked on the location-aware influence maximization. In their problem, each user has geographical location and the target users should be located in a given region. The paper proposed two greedy algorithms with 1-1/e approximation ratio and another two algorithms to meet the instant-speed requirement. The existing work considering location information in influence maximization only regards location as a simple user property and has not analyzed the user mobility behaviors. But in the real O2O environment, the influence spreading is closely related with the user location preference. Analysis on the user historical behaviors should be taken and the problem based on the mobility behavior is worthy to be solved.

The existing researches mainly concern about the single aspect of online social network and regard location as a simple property. It is needed to consider the influence maximization problem which describe the real O2O promotion and makes full use of location information. How to model the real problem in O2O environment and design the solution for the proposed problem is the important content of this paper.

3. DATASET

A significant feature and novelty of our research is to describe the objective fact, so we would like to carry out our research on the basis of real data.

LBSN is an important platform for influence diffusion in O2O environment. It contains both user social network and geographic information which could provide conditions for our research. We conduct our research on the dataset of Foursquare, the biggest and the most popular LBSN at present.

As our problem is associated with both the online and offline information, we need to retrieve both the user network and user check-ins at the same time. Because of the API restriction, the user historical check-ins could not be crawled directly. To solve the problem, we make use of the interaction of Foursquare and Twitter. Since users of Foursquare could bind their Twitter account and share their check-ins as tweets on Twitter, we could retrieve the historical tweets of a user and further manipulate the tweets to get the historical check-ins. By this way, we obtain the complete dataset consisting of online social network and offline historical check-ins.

The final dataset we use includes users living in New York and check-ins made in New York, as only the check-ins occurred in the living city of a user could be the reference to determine his or her daily geographic preference. Meanwhile, to a seller who provides life service product, people in the same city would be the main target that might bring long-term and stable benefit, while people out of the same city must be restricted by space and time to conduct the consuming behavior for the product. We extract the New York dataset based on the property of the users and check-ins. In the other aspect, we only concentrate on the consuming check-ins as we are interested in the geographic preference of user consuming behavior. Therefore the data is further filtered according to the category information offered by Foursquare to get the consuming check-in dataset

Table 1: Dataset parameters

Users	10901
Edges	170048
Average degree	15.599
Network diameter	9
Average path length	3.761
Check-ins	764328
Average check-ins	70.115

in New York which contains a variety of consuming behavior types. The basic parameters of the final dataset are listed in Table 1.

4. TWO PHASE MODEL AND LOCATION-BASED INFLUENCE MAXIMIZATION

4.1 Two Phase Model

The influence diffusion process under the O2O environment contains transition between online behavior and offline behavior. When users are influenced by others online, they would take the offline experience first before determining whether to accept the product or not. However, in traditional diffusion model such as Linear Threshold Model (LT) [11] and Independent Cascade Model (IC)[12, 13], only the online diffusion process exists while the offline part is always missing. So a new influence diffusion model is needed to describe the product acceptance considering both online and offline sections.

Whether a user would take an offline experience after being influenced online lies on the location property of both the user and the promoted product, because the consuming behavior of a user is affected by his or her consuming location preference which reflects the daily activity area. The further the product is away from the user daily activity area, the lower the probability is for the user to try out the product offline. Thus for a given product location L, each user has a probability to consume there, which is also the probability to accept the product offline. To introduce the new diffusion model, we first give the following definitions.

Definition 1. **Online Probability** $p^{on}_{i,j}$ The probability that user i successfully influences user j through edge $< i, j >$, which corresponds to the probability in the traditional influence diffusion model. This is a property on edge $< i, j >$.

Definition 2. **Offline probability** $p^{off}_{i,L}$ The probability that user i goes to location L for consuming, which is the probability that user accepts the product offline. This is a property of node i.

There are two phases in our diffusion model: online phase and offline phase, which reflects the process of accepting a product under the O2O circumstances. The model defines four states of a user:

Definition 3. **Inactive state** In this state, the user has not been online influenced by the neighbor.

Definition 4. **Online-active state** In this state, the user has been influenced by the neighbor and accepted the product online, but the offline experience has not been conducted.

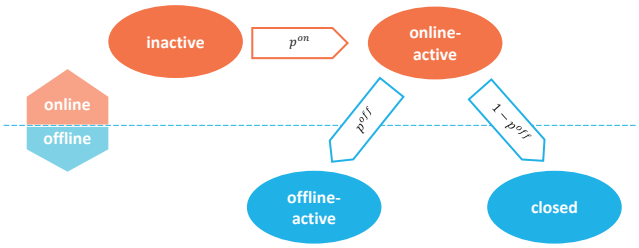


Figure 1: User state transition diagram

Definition 5. **Offline-active state** After transiting into the online-active state, the user steps into the offline phase. The offline-active state means the user has accepted the product offline.

Definition 6. **Closed state** After stepping into the offline phase, the user does not accept the product and moves into the closed state. Thereafter the user could not be influenced anymore and will never transit into any active states.

The user state transition diagram is shown in Fig. 1.

We propose the **Two Phase Model (TP model)** on the basis of the IC model, describing the process of product promotion involving the geographic property. Given a product to be promoted locating in L, and an initial node set A_0 in which the users has already been in the offline-active state, the procedure of the influence diffusion in TP model unfolds in the following discrete steps:

In step t, if node v first becomes offline-active, it will have a chance to activate each currently inactive neighbor w to make w transit into online-active state. The activate attempt succeeds with the online probability $p^{on}_{v,w}$. If v succeeds, w will become online-active in step $t+1$, and then enter into the offline phase. In the offline phase, w may transit into offline-active state with the offline probability $p^{off}_{w,L}$; otherwise it will switch to closed state. No matter whether the activate attempt of v succeeds or not, v cannot try to activate w in the following rounds. If w has more than one neighbor that transit into online-active state in step t, then these neighbors could try to activate w in arbitrary order, but all the attempts must happen in step t. The diffusion process is terminated when there are no more nodes that could get into the offline-active state.

Connections exist between the two phases and four states in TP model. Inactive state and online-active state belong to the online phase, while offline active state and closed state belong to offline phase. The transition from online phase to offline phase starts at the online-active state, which means the user starts an offline experience, and in turn the offline-active state brings the offline behavior back to the online diffusion, which means the user gets back online and continues to activate other users.

In TP model, the offline probability $p^{off}_{w,L}$ is related to the product location L and the consuming location preference of user w, and we will discuss the details in Section 5. On the other hand, the online probability $p^{on}_{v,w}$ could be set or calculated according to the manner of traditional online model, e.g. set the equal value for all the edges or calculate the different probability of each edge according to certain rules. This separation of online and offline probabilities brings relatively high universality and flexibility to TP model.

4.2 Location-based influence maximization

The location-based influence maximization problem under the O2O circumstances is different from the traditional one in multiple aspects. Firstly, the traditional models are not proper anymore and TP model which aims at the O2O model should be adopted instead. Moreover, the target of influence maximization is more specific. As different users have different offline probability to carry out the consuming behavior for the same product, the target users of the promotion should be those who are more likely to conduct the actual consuming behavior. Thus the purpose of this new problem is to have the most target users influenced when the diffusion completes.

The **location-based influence maximization problem** could be formally described as following:

Given a directed graph $G(V, E)$, in which V is the node set representing the users in the social networks, and E is the edge set representing the relations between users. $\forall v_i \in V(1 \leqslant i \leqslant |V|)$, C_i denotes its consuming check-in set, and given a location $L = (lat, lon)$, v_i has the offline probability $p^{off}_{v_i,L}$ representing the probability with which v_i will consume in location L.

Given a TP Model, the initial nodes number k and the product location $L = (lat, lon)$, the location-based influence maximization problem is to find a set of initial nodes $S \subseteq V$, $|S| = k$, and after the diffusion completes according to the TP model, the final influenced target nodes could be the most. Here, the target nodes satisfies $p^{off}_{v_i,L} > \theta$, where θ is a given threshold.

The threshold θ could be set according to the real situation. It reflects the expectation of the promoter on the target users. Targeting on the users who have the higher possibility to consume could create more promotion benefits, because those users are more likely to become repeat customers. Even though the users out of the target set may have nonzero offline probability, obviously their activity areas are too far that they could not bring the long-term and stable benefit. This consideration also emphasizes the locality, which is an important feature of O2O model.

5. OFFLINE PROBABILITY

Offline probability is an important parameter to describe the offline behavior of a user. It is the probability for a user to head for a given location and consume there, which is essentially a representation of the user mobility pattern. The existing researches have revealed that the probability with which a user moves from one location to another has certain relation with the distance between the two locations. In the other aspect, the historical check-ins could be the real and explicit reflection of user mobility behavior. A single check-in shows a user appears in a certain location, while multiple check-ins could indicate the daily activity area of the user, making the historical consuming check-in record useful in the offline probability calculation. This section will combine the mobility pattern of single movement and the historical check-in record to propose calculation method of the offline probability.

5.1 User mobility pattern analysis

The displacement behavior of human beings follows certain pattern. Existing researches have discussed the relationship between displacement distance and probability.

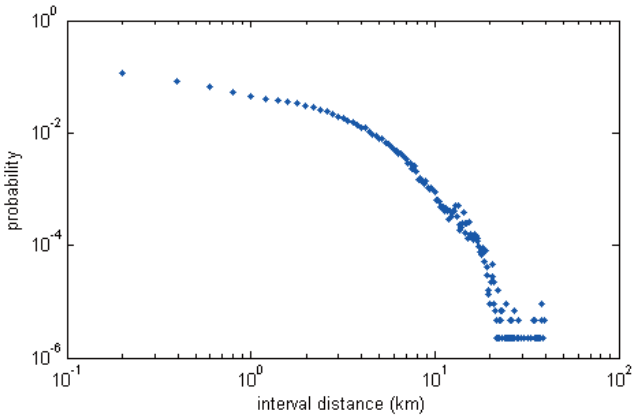


Figure 2: Displacement distribution of NY users

Some of them studied the behavior of animals and claimed that the trajectory of animals could be estimated by Lévy flight [14, 15]. In [16], authors looked into the mobility data collected from cellphones which could indicate the human trajectory and analyzed human mobility pattern. They found that the distribution of displacement distance could be approximated by a truncated power-law:

$$P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} exp(-\frac{\Delta r}{k}) \quad (1)$$

where Δr denotes the displacement distance of a user, Δr_0 and β are the parameters in power-law, and k is the cutoff value.

Inspired by the distribution, we conduct the statistics on the displacement distribution of New York users based on the Foursquare check-in dataset. We sort the check-in set C_u of user u according to the check-in time and get the check-in sequence $S_u = (c_1, c_2, ..., c_i)$. A single displacement distance d_i could be calculated as the distance of two adjacent check-ins from the sequence, which is $|c_i - c_{i-1}|$. We calculate the displacement distance of New York users and draw the displacement distribution as shown in Fig. 2.

As Fig. 2 indicates, the displacement distribution of New York users could also be approximated by a truncated power-law described in Eq.1. We use Eq.1 to fit the displacement data of New York users and get the parameter value in the equation. For users in New York, $\Delta r_0 = 1.69856$, $\beta = 2.41922$ and $k = 7.05365$.

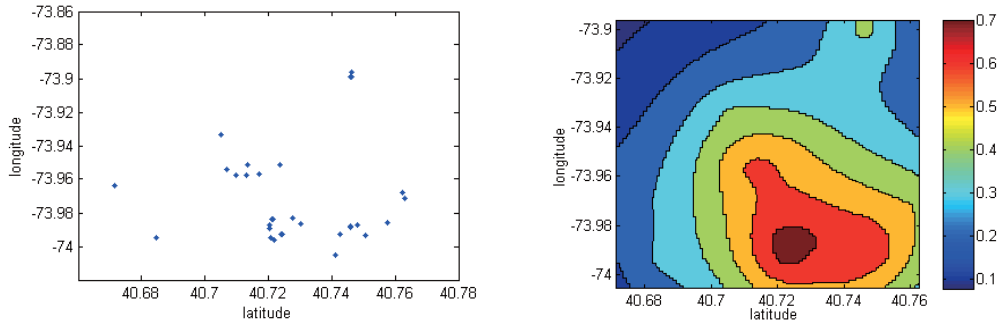
5.2 Calculation of offline probability

Based on the user mobility pattern, we could calculate the offline probability of a user utilizing the historical check-in record. For a single user in New York whose consuming check-in set is C, given a product location L, the offline probability with which the user consumes in location L after it transits into online-active state could be calculated as:

$$P^{off}_L = 1 - \prod_{all\ c_i \in C} (1 - p_i) \quad (2)$$

$$p_i = (dis_i + \Delta dis)^{-\beta} exp(-\frac{dis_i}{k}) \quad (3)$$

where dis_i denotes the distance from historical consuming check-in c_i to location L, $\Delta dis = 1.69856$, $\beta = 2.41922$ and



(a) Consuming check-in of a user

(b) Contour map of its offline probability

Figure 3: Example of offline probability

$k = 7.05365$. Eq.3 comes from Eq.1. p_i means for each historical consuming check-in c_i , the probability with which the user leaves from the position of c_i to location L for consuming.

The consuming check-in of a user is the reflection of his or her daily activity area. If a user has a check-in in location A, it implies the user appears in location A so that A is a possible activity position of the user. So we could assume the user may start from location A and head to location L for consuming and further assume each historical consuming check-in location is a possible starting point for the consuming behavior.

In Eq.2, the physical meaning of $\prod_{all\ c_i \in C}(1 - p_i)$ is the probability that no matter which historical consuming check-in location the user departs from , he or she will not conduct any consuming behavior in location L. Thus the probability that this situation does not occur will be P_L^{off} , which could be calculated by Eq.2. Its physical meaning is that departing from any of the historical consuming check-in locations, the user would get to location L and conduct consuming behavior for at least one time, which is actually the offline probability in our problem.

Fig. 3 (a) plots the consuming check-ins of a randomly selected user in the longitude and latitude coordinate system. Fig. 3 (b) is the contour map of its offline probability. Fig. 3 shows the consistency of offline probability distribution and historical consuming behavior. The distribution tends to spread from several centers. The central parts with higher probability correspond to the areas with denser check-ins, which are actually the daily activity centers of the user. Obviously, it is relatively more possible for the user to have consuming behavior in these areas. In another aspect, the further the location is apart from the center, which means it has fewer check-ins around, the lower the probability is for the user to go consuming there. So in general the distribution is consistent with the practical situation. What should be emphasized is that although only a few check-in locations appear in the upper-right part, the check-in times in these locations are relatively more, so a small center is formed in the contour map there, showing that it is still an activity area of the user. This proves the comprehensiveness of the calculation method on reflecting the consuming location preference.

Based on Eq.2 and Eq.3, given any product location L, for each user i in the social network, the offline probability could be calculated and be utilized in the TP model and location-based influence maximization.

6. LOCATION-BASED INFLUENCE MAXIMIZATION ALGORITHM AND EXPERIMENT EVALUATION

6.1 Two-Phase Heuristic Algorithm

In this section, we would propose a heuristic algorithm to solve the location-based influence maximization problem. In TP model, each node has its own offline probability for a given product, so the consideration on the location-based influence maximization cannot only focus on the network topology, but also the offline property of each node. We first propose a heuristic parameter H based on TP model, which takes both network topology and online-offline probability into account, to measure the influence of each node.

Given a location L, the value of H could be calculated with Eq.4.

$$H_i = \sum_{each\ neighbor\ j\ of\ i} p_{i,j}^{on} p_{j,L}^{off} \tag{4}$$

Here, $p_{i,j}^{on}$ is the probability that user i influences user j though edge $< i, j >$, namely the online probability on edge $< i, j >$, while $p_{j,L}^{off}$ is the probability that user j has consuming behavior in location L, namely the offline probability of user j.

The definition of H combines the influence propagation ability and propagation range of a user at the same time. For a user i who has a neighbor j, $p_{i,j}^{on} p_{j,L}^{off}$ describes the ability user i has to push user j into offline-active state. The larger the value is, the more possible it is for the influence to propagate from user i to user j. So this value could outline the propagation ability. Adding the values of all the neighbors together, in the other aspect, considers the network topology. A user with more neighbors has more chances to influence others. So the addition process outlines the propagation range. In general, the heuristic parameter H could serve as a synthesized indicator to measure the user influence.

Based on the heuristic parameter, we propose the Two-Phase Heuristic Algorithm (TPH algorithm) to solve the location-based influence maximization problem. TPH algorithm calculates the value of H_i for each user node i, and

select k nodes with largest H value as initial nodes. It is summarized in Algorithm 1.

Algorithm 1 Two-Phase Heuristic Algorithm (TPH)

Input: Social network $G(V, E)$, TP model, initial nodes number k, product location L
Output: initial nodes set S (with the size of k)
1: initialize $S = \varnothing$;
2: calculate the value of heuristic parameter H for all of the nodes in V;
3: **for** $i = 1$ to k **do**
4: select node u in $V \backslash S$ with the largest H value;
5: $S = S \cup \{u\}$;
6: **end for**
7: **return** S;

For a given product location L, each node in social network G has a corresponding offline probability.

In Algorithm 1, line 2 calculates the value of H for each node according to Eq.4. Line 3 to line 6 select k nodes with largest H value and add them into the initial node set. As the calculation of H for each node needs to process all of its neighbors, the time complexity of line 2 is $O(m)$, in which m denotes the number of edges in the network. The complexity of the selection procedure from line 3 to line 6 is $O(klogn)$, where n denotes the number of nodes in the network. So the total time complexity of TPH algorithm is $O(klogn + m)$.

6.2 Experiment Evaluation of TPH

In order to evaluate the performance of TPH Algorithm, this section compares it with several common heuristic algorithms. The experiment runs in the PC machine with Intel Core i5 2.80GHz processor, 2G memory and 64 bit Windows 7 operating system. The experiment program is coded in C++.

As explained in Section 4, TP model offers more accurate description for influence diffusion in O2O environment. The experiment will execute all the algorithms based on TP model to prove the effectiveness of TPH algorithm on location-based influence maximization problem. According to the definition of the problem in Section 4, the evaluation criterion of the experiment is the number of the influenced target users. The algorithm that obtains more influenced target users has the better performance in influence diffusion.

6.2.1 Experiment setup

The dataset adopted in the experiment is the real one in Foursquare introduced in Section 3, consisting of the user network in New York and the historical consuming check-ins located in New York. Considering the compatibility with our problem, we choose three common and representative heuristic algorithms to make comparison: Max Degree algorithm (Degree), Degree Discount algorithm (DD) and PageRank algorithm (PR). Max Degree algorithm selects k nodes with largest degree as initial seeds. DD algorithm makes certain discount on the neighbor degree of selected seed node. Here, we use the online probability in the discount calculation. PR algorithm is the core search algorithm of Google and could be used in ranking process. Here, we set the damping ratio in PR algorithm as 0.85. Besides, random algorithm is executed as a reference.

Table 2: Experiment parameters

Product location L	Popular	Normal	Unpopular
Target user threshold θ	0.3	0.5	
Online probability p_{on}	0.01	0.05	0.1

When we set the location-based influence maximization problem, the product location L and the target user threshold θ should be given. A single user will have different offline probability for different location L and different threshold would determine different target user group. So in order to provide a comprehensive evaluation on the algorithm performance, the experiment generates specific influence maximization problems with different L and threshold. Besides, as the experiment set the same online probability p_{on} for each edge in TP model according to the traditional manner and the value of p_{on} will affect the diffusion result, different online probability would be set to make an exhaustive analysis. The choices of the three parameters L, θ and p_{on} are shown in Table 2.

We have defined three location types based on the popularity. People are more likely to consume in the popular locations but much less in the unpopular location. Normal locations are the non-extreme ones between popular and unpopular locations. According to the definition, it is obvious that the general offline probabilities of the user group to consume in a popular location are relatively high, while those to consume in an unpopular location are relatively low. Thus, the average offline probability of all the users, denoting as $AVG(p_{off})$, could be utilized as the criterion to classify the locations. We analyze a large number of samples and set the following rules:

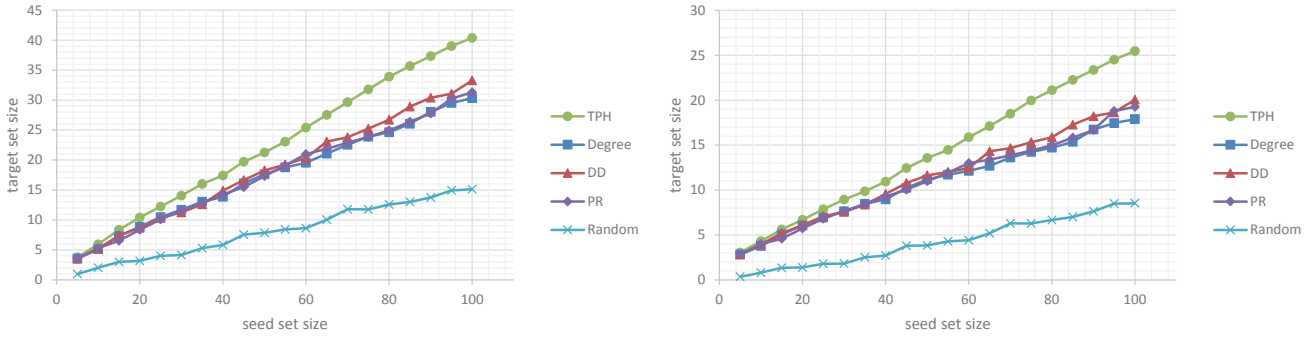
$$location\ L\ is \begin{cases} popular & if\ AVG(p_{off}) \geqslant 0.4 \\ normal & if\ 0.04 \leqslant AVG(p_{off}) < 0.4 \\ unpopular & if\ AVG(p_{off}) < 0.04 \end{cases}$$

To avoid overfitting, we randomly select 10 locations for each location type, and the final experiment result is the average of the 10 cases for each configuration setting.

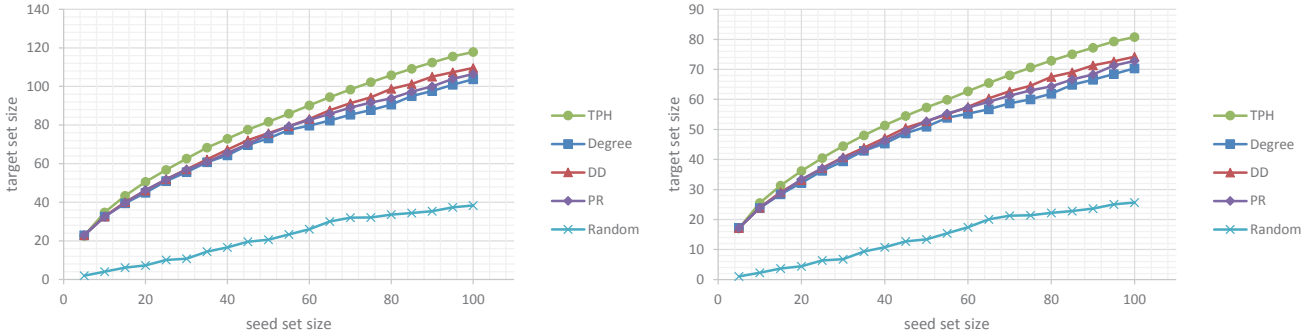
The target user threshold θ is set to 0.3 or 0.5, respectively representing a lower and a higher expectation on the target user. If θ is higher, the target users would be extremely few when L is not popular, which is not suitable for experiment analysis. Moreover, the promoters in real life would not set excessively harsh criteria for target user. So the higher threshold here is 0.5. The online probability p_{on} is set to 0.01, 0.05 or 0.1. If the offline probabilities are generally low, then a relatively high online probability would be chosen to make the result clear.

6.2.2 Experiment result and analysis

When L is a normal location, the experiment result is illustrated in Fig. 4. The horizontal axis shows the number of initial nodes and the vertical axis shows the number of influenced target nodes after the diffusion terminates. Fig. 4 illustrates significant superiority of TPH algorithm over the others when L locates in normal area. The three compared algorithms have similar performance among which the DD algorithm performs a little better. The values of online probability and threshold have effect on the final number of influenced target users, but they do not change the overall trend of the experiment data. TPH algorithm outperforms

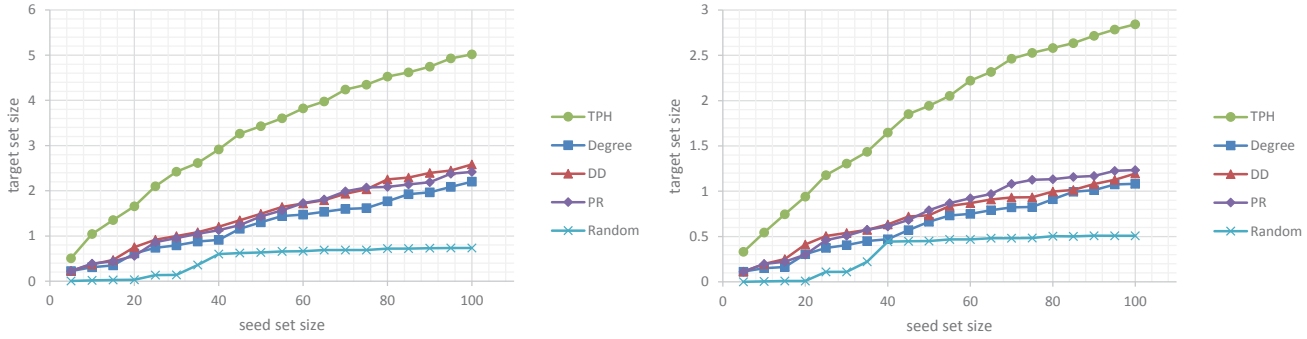


(a) p_{on}=0.01 threshold=0.3 (b) p_{on}=0.01 threshold=0.5

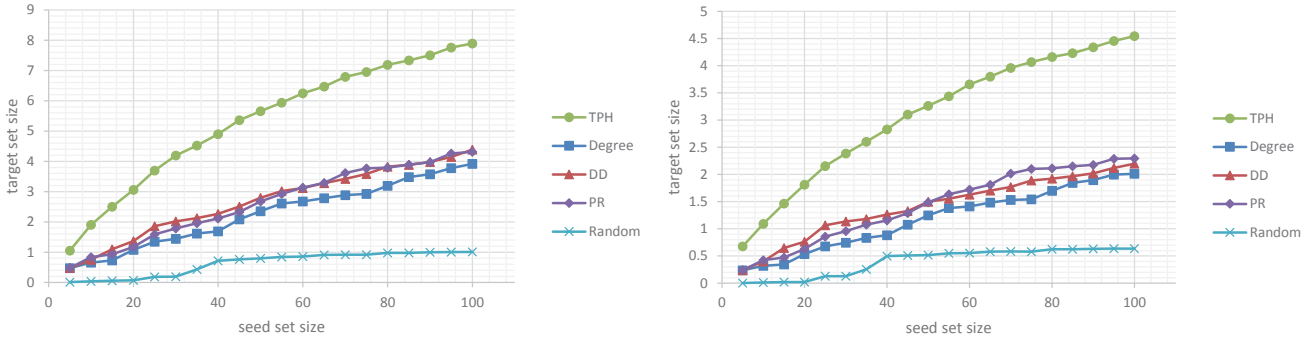


(c) p_{on}=0.05 threshold=0.3 (d) p_{on}=0.05 threshold=0.5

Figure 4: Performance of different algorithm when L a normal location

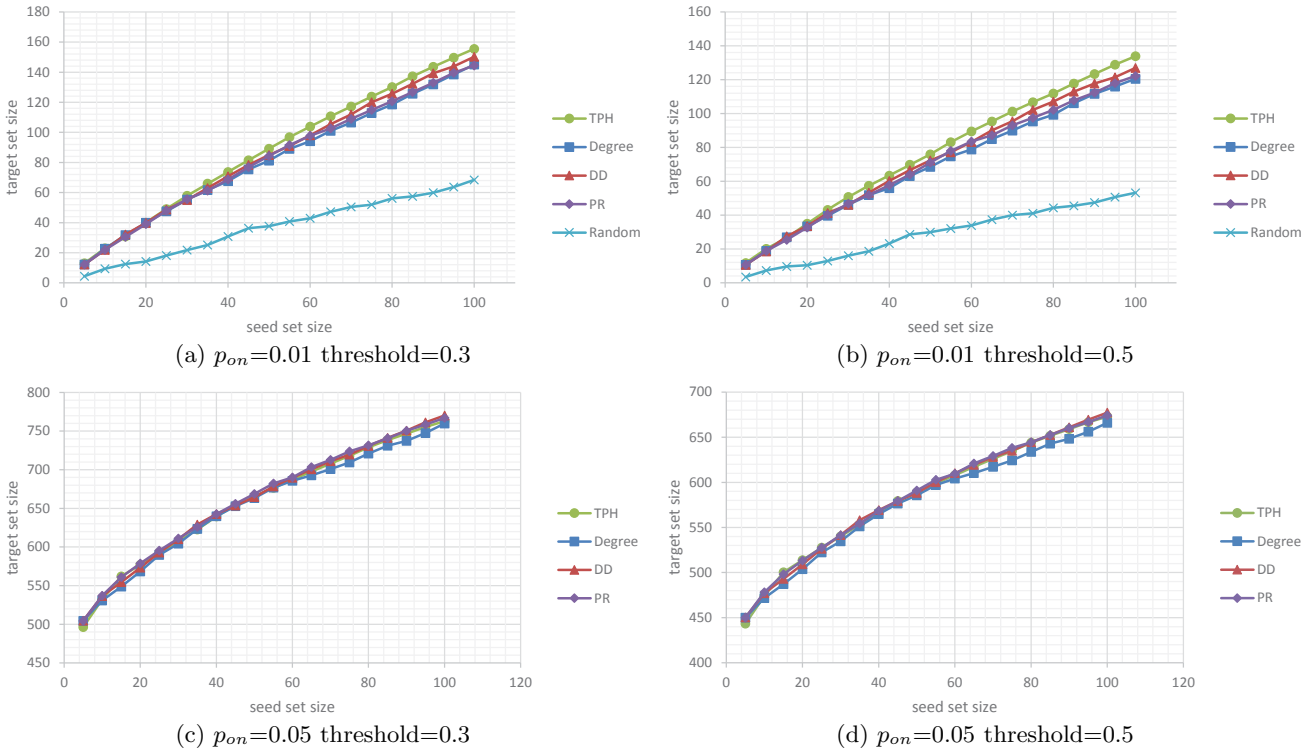


(a) p_{on}=0.05 threshold=0.3 (b) p_{on}=0.05 threshold=0.5



(c) p_{on}=0.1 threshold=0.3 (d) p_{on}=0.1 threshold=0.5

Figure 5: Performance of different algorithm when L is an unpopular location



(a) p_{on}=0.01 threshold=0.3 (b) p_{on}=0.01 threshold=0.5

(c) p_{on}=0.05 threshold=0.3 (d) p_{on}=0.05 threshold=0.5

Figure 6: Performance of different algorithm when L is a popular location

the other algorithms in all cases when online probability and threshold are different.

When L is an unpopular location, the experiment result is shown in Fig. 5. It again shows the better performance of TPH algorithm. Similarly, the change of online probability and threshold has no significant influence on the trend of experiment data and TPH algorithm performs better in each case.

When L is a popular location, the experiment result is shown in Fig. 6. The performance of random algorithm is far behind the others while others have similar results in Figs. 6(c) (d), so we do not illustrate the curve of random algorithm in the graph. Fig. 6 indicates a relatively big effect of the online probability on the diffusion result when L is popular. When online probability is 0.01, TPH algorithm again shows superiority in diffusion result and the performance of Degree Discount algorithm is close to TPH. However, when the online probability comes to 0.05, TPH loses its advantage and all the algorithms have similar diffusion results and increasing trend. This phenomenon is related to both online and offline probability. As L being popular, the offline probabilities of most users will be generally high. When the online probability is further set to a relatively high value, the overall influence diffusion ability of the network becomes strong and a large number of users could be influenced. Each algorithm tries to choose the nodes with higher influence indicator it defines, but when the diffusion ability of the network itself is already quite strong, the difference between different selection methods is relatively slight. It is because the seed nodes chosen by one algorithm may be covered during the diffusion process of another algorithm, which will decrease the dissimilarity of individual algorithms. Unlike online probability, threshold has little effect on the general diffusion trend.

6.2.3 *Experiment conclusion*

Among the three experiment parameters, the effect of target user threshold is insignificant on the performance trend of the algorithms. The product location L significantly affects the experiment result and when L is a popular location, the online probability also has influence on the algorithm performance.

In general, TPH algorithm has outstanding advantage over the other ones. When L is not popular, this advantage is more significant, while in the special case where L is popular and online probability is high, the performance of TPH algorithm is similar with other ones.

6.3 Multi-Rank Algorithm

TPH algorithm is not satisfying when product location L is popular and online probability is relatively high. In order to make a complement, we propose the Multi-Rank algorithm (MR algorithm) which could improve the diffusion result in this situation.

The main idea of MR algorithm is to take the activation accessibility into account, which reflects whether a user is easy to be activated (to transit into offline-active state). The algorithm will select the nodes which are more likely to influence others from those with low activation accessibility.

Looking into the initial nodes selected by TPH algorithm in the last case of the experiment, we notice that as the influence diffusion ability of the network is strong, there is high possibility for those selected nodes to have high activation accessibility, which means even if we do not select them as initial nodes, the influence may spread to them

later. However, once the influence propagates to the nodes with low activation accessibility, the diffusion process may be terminated and the influence could not cover the other areas which are connected by these nodes, making the diffusion performance stuck. So we consider the problem in a brand new way, selecting the initial nodes from those with low activation accessibility, which could activate the areas that are formerly difficult to be covered. Moreover, as the general diffusion ability is strong, some nodes with high activation accessibility which are selected by TPH but not by MR could also be covered in the diffusion process.

When selecting from the nodes with low activation accessibility, we also apply the idea of TPH algorithm. The heuristic parameter H in TPH algorithm is adopted to guarantee the influence diffusion ability carried by the initial nodes themselves is relatively high.

In order to measure the activation accessibility, we first define a ranking basis R. It is utilized to rank the node to distinguish whether it is easy to transit into offline-active state. Given a user i and product location L, the value of R_i could be calculated as:

$$R_i = R_i^{neighbor} + R_i^{self} \tag{5}$$

$$R_i^{neighbor} = 1 - \prod_{each\ neighbor\ j\ of\ i} (1 - p_{j,L}^{off}) \tag{6}$$

$$R_i^{self} = p_{i,L}^{off} \tag{7}$$

The larger R_i is, the easier user i is to be activated. When we say a user has high activation accessibility, first of all, he or she should be easy to switch to offline-active state, and meanwhile the neighbors of this user should also be easy to be activated. If none of the neighbors are activated, the user could never be activated at all. So the definition of R_i has two parts, in which $R_i^{neighbor}$ denotes the activation accessibility of neighbors of user node i and R_i^{self} denotes the activation accessibility of user node i itself. With regard to $R_i^{neighbor}$, $p_{j,L}^{off}$ in Eq.6 is the offline probability of user j. The physical meaning of multiplying all the $(1 - p_{j,L}^{off})$ is to calculate the probability that none of the neighbors of user node i would transit into offline-active state. So the physical meaning of $R_i^{neighbor}$ is the probability that at least one of these neighbors become offline-active from online-active state. The higher $R_i^{neighbor}$ is, the higher activation accessibility the neighbors have in general. For R_i^{self} , its value is the offline probability of user node i. The higher R_i^{self} is, the higher activation accessibility i itself has. The value of R_i is the sum of $R_i^{neighbor}$ and R_i^{self} .

MR algorithm first calculates the ranking basis R for all the nodes, and then ranks each node according to the value of R. The method of ranking is to sort the nodes by a descending order of R, and the former half are defined as high-ranking nodes while the latter half are defined as low-ranking nodes. High-ranking nodes are those with high activation accessibility and low-ranking nodes are the opposite. At last, MR algorithm selects k nodes with largest influence from the low-ranking node set as the initial nodes applying TPH algorithm. Its process is summarized in Algorithm 2.

As the same as TPH algorithm, for a given product location L, each node in social network G has a corresponding offline probability. In Algorithm 2, line 2 calculates the value of R for each node according to Eq.5. Line 3 to line 6 rank the nodes and create the low-ranking node set. Line 7 to line 11 apply TPH algorithm to choose the initial nodes from the low-ranking node set. The time complexity of calculating the value of R and H is $O(m)$ and the complexity of ranking user nodes is $O(nlogn)$, in which m and n denote to the number of edges and nodes in the network respectively. The complexity of selecting k initial nodes is $O(klogn)$. So the total time complexity of MR algorithm is $O(nlogn+m)$.

Algorithm 2 Multi-Rank Algorithm (MR)

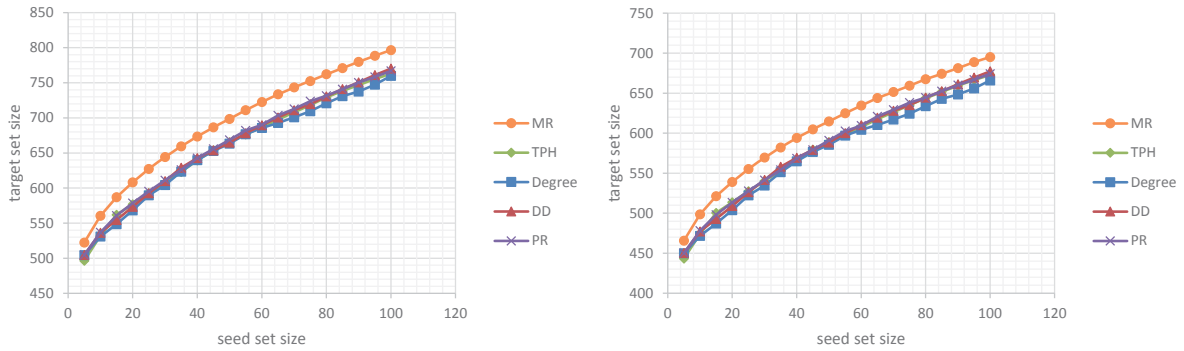
Input: Social network $G(V, E)$, TP model, initial nodes number k, product location L
Output: initial nodes set S (with the size of k)
1: initialize $S = \varnothing$, $LS = \varnothing$; // LS is the set of low-ranking nodes
2: calculate the ranking basis R for all the nodes in V;
3: sort the nodes in V by the descending order of R and obtain the node sequence $(v_1, v_2, ..., v_n)$;
4: **for** $i = n/2$ to n **do**
5: $LS = LS \cup \{v_i\}$;
6: **end for**
7: calculate the value of heuristic parameter H for all of the nodes in LS;
8: **for** $j = 1$ to k **do**
9: select node u in $LS \backslash S$ with the largest H value;
10: $S = S \cup \{u\}$;
11: **end for**
12: **return** S;

6.4 Experiment Evaluation of MR

In order to prove the effectiveness of MR algorithm, we run the comparison experiment together with the former algorithms. As MR algorithm is a complement for TPH algorithm when product location L is popular and online probability is high, we set the experiment according to the third case in section 6.2.2, where L is popular and the online probability is 0.05. Again, 10 popular locations are randomly selected to calculate the average data. The result in Fig. 7 shows the outstanding performance of MR algorithm and also proves its steady advantage. While other algorithms have similar diffusion results, MR algorithm brings a large improvement in the range of diffusion. MR algorithm is proved to be an effective complement for TPH algorithm when L is popular and online probability is high.

7. CONCLUSION

This paper aims at the product promotion in O2O model and carries out the study of location-based influence maximization on the platform of LBSN. We first propose the Two Phase Model, a new diffusion model which could be adopted under the O2O environment. Then the location-based influence maximization problem is defined and the calculation method of offline probability is present by analyzing the user mobility pattern. Next, the TPH algorithm is proposed for location-based influence maximization and MR algorithm is designed as the complement. At last, experiments prove TPH and MR algorithm has significant effectiveness. In the future work, we would further consider the semantic information of consuming location and analyze user location preference of different semantic categories. Moreover, we would introduce the temporal property of user consuming behavior



(a) p_{on}=0.05 threshold=0.3 (b) p_{on}=0.05 threshold=0.5

Figure 7: Performance of MR algorithm

to study the influence diffusion with time restriction so that the more general solution could be raised.

8. ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (61272531, 61202449, 61272054, 61370207, 61370208, 61300024, 61320106007 and 61472081), China high technology 863 program (2013AA013503), Jiangsu Technology Planning Program (SBY2014021039-10), Jiangsu Provincial Key Laboratory of Network and Information Security under Grant No.BM2003201 and Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grant No.93k-9.

9. REFERENCES

[1] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001.

[2] Yu Zheng and Xiaofang Zhou. *Computing with spatial trajectories*. Springer Science & Business Media, 2011.

[3] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[4] David JC MacKay. Introduction to monte carlo methods. In *Learning in graphical models*, pages 175–204. Springer, 1998.

[5] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429. ACM, 2007.

[6] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. Celf++: optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th international conference companion on World wide web*, pages 47–48. ACM, 2011.

[7] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009.

[8] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[9] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[10] Guoliang Li, Shuo Chen, Jianhua Feng, Kian-lee Tan, and Wen-syan Li. Efficient location-aware influence maximization. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 87–98. ACM, 2014.

[11] H Peyton Young. The diffusion of innovations in social networks. *The Economy As an Evolving Complex System III: Current Perspectives and Future Directions*, 267, 2006.

[12] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

[13] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[14] Joseph Klafter, Michael F Shlesinger, and Gert Zumofen. Beyond brownian motion. *Physics today*, 49(2):33–39, 1996.

[15] Rosario N Mantegna and H Eugene Stanley. Stochastic process with ultraslow convergence to a gaussian: the truncated lévy flight. *Physical Review Letters*, 73(22):2946–2949, 1994.

[16] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.