To Return or to Explore: Modelling Human Mobility and Dynamics in Cyberspace

Tianran Hu
University of Rochester
thu@cs.rochester.edu

Yinglong Xia
FutureWei Technologies, Inc.
yinglong.xia.2010@ieee.org

Jiebo Luo
University of Rochester
jluo@cs.rochester.edu

ABSTRACT

With the wide adoption of multi-community structure in many popular online platforms, human mobility across online communities has drawn increasing attention from both academia and industry. In this work, we study the statistical patterns that characterize human movements in cyberspace. Inspired by previous work on human mobility in physical space, we decompose human online activities into ***return*** and ***exploration*** – two complementary types of movements. We then study how people perform these two movements, respectively. We first propose ***a preferential return model*** that uncovers the preferential properties of people returning to multiple online communities. Interestingly, this model echos the previous findings on human mobility in physical space. We then present ***a preferential exploration model*** that characterizes exploration movements from a novel online community-group perspective. Our experiments quantitatively reveal the patterns of people exploring new communities, which share striking similarities with online return movements in terms of underlying principles. By combining the mechanisms of both return and exploration together, we are able to obtain an overall model that characterizes human mobility patterns in cyberspace at the individual level. We further investigate human online activities using our models, and discover valuable insights on the mobility patterns across online communities. Our models explain the empirically observed human online movement trajectories remarkably well, and more importantly, sheds better light on the understanding of human cyberspace dynamics.

CCS CONCEPTS

• **Information systems** → **Social networking sites**; *World Wide Web*; • **Applied computing** → **Sociology**.

KEYWORDS

Cyberspace, Online Communities, Human Mobility, Human Dynamics, Preferential Return, Preferential Exploration

ACM Reference Format:
Tianran Hu, Yinglong Xia, and Jiebo Luo. 2019. To Return or to Explore: Modelling Human Mobility and Dynamics in Cyberspace. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3308558.3313686

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313686

1 INTRODUCTION

In physical space, human movements follow a set of underlying principles. Much research work has been devoted to human mobility in physical space on various topics such as individual and group level mobility patterns [11, 39], the temporal-spatial properties of human mobility [17], and the relation between individual mobility and social connections [9]. With the rapid advances in information and communication technologies (ICT), especially the smartphones, the majority of the population in modern societies around the world are spending a significant and increasing amount of their daily time and activities in cyberspace, and in doing so connecting to multiple online communities.

Uncovering the statistical patterns that characterize the trajectories of human movements across online communities is not only key to many research areas such as computational sociology [4, 51], social media [19, 35], and complex networks [36], but also of great importance for the design of multi-community platforms [20]. Much work has been done on modelling online trajectories, including learning sequential patterns from user traces [31], detecting information spreading among communities [21, 43], and extracting human mobility patterns in cyberspace [18]. However, these studies are usually descriptive and limited at a coarse user-group granularity, while a finer grained understanding in the underlying rules that people follow when they move across online communities is still missing. To address the deficiency, we employ a popular framework in physics for studying individual movements across real word locations [32, 40]. To be specific, we decompose online activities to two basic types of movements across communities – ***return*** and ***exploration***. We then study based on what mechanisms individuals perform these two types of movements.

We start our work with investigating the relation between return and exploration. When an individual performs a move in cyberspace, she either returns to a previously visited online community, or explores a new community. Previous work on human online activities suggests that, at each move, users are not choosing between return and exploration randomly [2]. One evidence is that the probability of people visiting new communities shows a decreasing tendency over time instead of staying constant [54], indicating an evolving relation between return and exploration. We take a close look at this evolving relation, and quantitatively show that the probability of exploring new communities decreases exponentially with the amount of unique communities previously visited. Generally speaking, this finding implies that the more communities an individual has visited, the more likely she will return to a visited community at her next move, rather than exploring a new one. Interestingly, this relation in cyberspace is similar to the relation between human physical return and exploration in real world, which is originally reported in the preferential return model in physics [40].

With the relation between return and exploration being modelled, we proceed to investigate how humans perform return movement among visited communities. Reportedly, people visit online communities following Zipf's law [7], i.e., an individual's preferred communities receive exponentially more visits than the less preferred. This observation indicates that the probabilities of returning are not evenly distributed among visited communities. Our experiments show that the probability of an individual returning to a community is proportional to the visit frequency she pays to this community. In other words, people are more intended to return to their frequently visited communities. Consequently, the frequently visited communities get even more returns (i.e., the rich get richer), and result in the exponential imbalance. This finding, again, echos the property of people returning to offline locations in physics preferential return model. Therefore, without causing confusion, we also refer to our model for online return movement (including its relation with exploration) as the ***preferential return model***.

We then study the mechanism how humans explore the cyberspace. Usually, people explore online communities driven by their personal demands and interests [12, 53]. This suggests that people are more likely to explore new communities somewhat related to the communities they have visited before, rather than randomly select a unvisited one. In fact, previous literature indicates that the online communities an individual visits can be categorized into ***community-groups*** [13, 29]. In other words, with a high probability, a newly visited community can be classified into one of the previously explored community-groups. Following this idea, we investigate the exploration movement from a community-group perspective. To be specific, given the group label of each online community as well as a user's exploration trajectory, we study how she selects new communities from her previously explored community-groups, or opens a new community-group to explore.

Our experiments quantitatively indicate that exploration movement at community-group level share a remarkable resemblance with return movement, and also follows a ***preferential principle***. We first focus on the probability that an individual explores a community in a community-group that she has never explored before. We discover that this probability also evolves over time, and decreases exponentially with the total amount of community-groups the user has explored. Next, we examine the probability of a user exploring a new community from her previously explored community-groups. Similar to return movement, our results show that people pay more attention to their preferred community-groups. To be specific, with much higher probabilities people select new communities from the preferred community-groups to explore. To the best of our knowledge, this is the first time that such a quantitative mechanism of human exploration in cyberspace being reported. We refer to the mechanism as the ***preferential exploration model***.

Through the lens of both *preferential return* and *preferential exploration* models, we are able to not only see a big picture of how people move across online communities, but also uncover more insights in human mobility patterns in cyberspace. For example, we uncover that some communities attract more people to return. The communities of this type are usually of relatively small sizes, and concentrate on niche topics. Meanwhile, we also find that the members of some other communities are more intended to explore. The communities of this type, differently, are usually of larger sizes, and their topics are broader and more interactive. In fact, our analysis indicates that more than 67% of the variance of user intention to explore new communities can be explained using user visit history. It is also of high research value to compare human dynamics in physical space and cyberspace. Interestingly, our experiments indicate that humans are more likely to explore in the physical world, while in cyberspace they are more likely to return to their familiar communities.

The main contributions of this paper are summarized as follows.

- We systematically study human mobility in cyberspace. In particular, we present a ***preferential return model*** and a ***preferential exploration model***. The models explain empirical evidences remarkably well, and synergistically characterize the statistical patterns of human movements across online communities.
- We further investigate online activities at an individual level using our models. Our findings reveal previously undiscovered insights on human cyberspace dynamics, such as the relations between user activities and the community features, the differences between human online and offline movements, and so on.

2 DATA COLLECTION AND PREPROCESSING

We collect our data from Reddit, one of the most visited websites in the world[1]. Reddit has been used as the data sources in many computational sociology studies [47, 50], due to its high popularity and almost complete data availability[2]. In addition, we select Reddit for our study for two other reasons. 1) Multi-community structure – Reddit is organized into thousands of communities (i.e. subreddits), and users are allowed to post in any communities at will. Such a multi-community setting breeds a large amount of activities across communities [15, 21, 51], and therefore makes the data source ideal for studying human online movements. 2) Complete user traces – all the user activities (except for the ones deleted by users) on the platform are available. In other words, up to a time point, we are able to collect complete user online trajectories from their first activities. Given that human online mobility patterns are evolving over time [16], using complete user traces for our study avoids possible information lose and leads to more precise analysis.

In this work, we take a user post in a Reddit community as a "check-in" activity in the community, and a movement between communities consists of two consecutive posts. Please note that the two consecutive posts can be located in the same community. We first download all the user posts on Reddit from 2012 July to 2014 Dec. This initial dataset contains totally 1 billion records, posted by 9 million users in 0.2 million online communities. We first filter out communities with less than 100 users, resulting in 17 thousand active communities. We then find users who did not join Reddit until 2013 Jan (i.e. remove all the user who have activities (posts) in the first six months in our data), and extract the trajectories of these users across active communities. By doing so, we ensure that the extracted trajectories are complete (i.e. starting from the users'

[1] en.wikipedia.org/wiki/Reddit
[2] Reddit data is publicly available, and free for download at www.reddit.com/3bxlg7

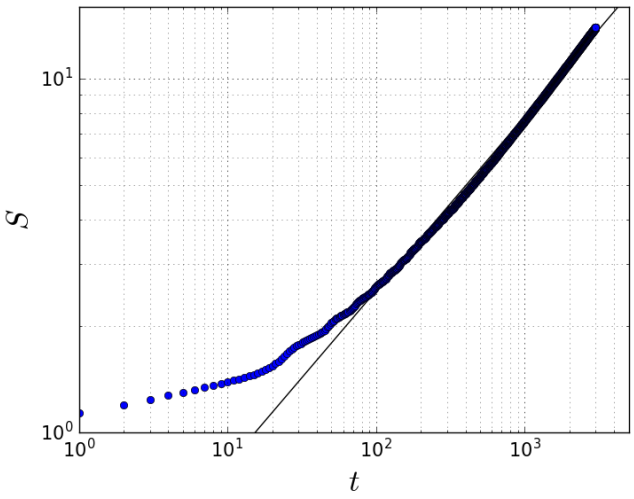

Figure 1: Distribution of the number of visited unique communities S over time t in hours. The fitted straight line indicates that $S \sim t^{\mu}$, where μ is estimated to be 0.42.

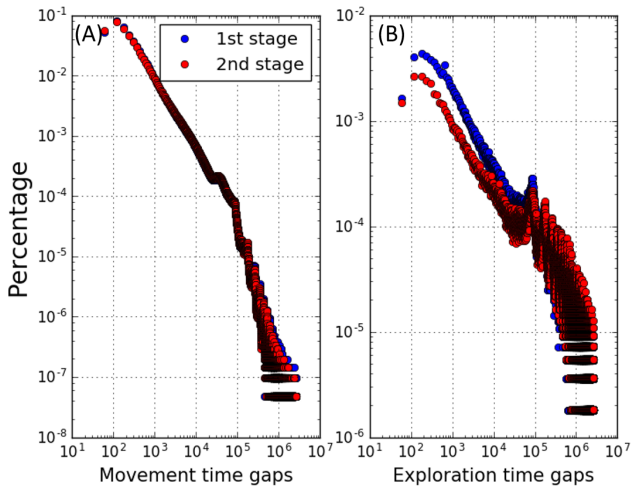

Figure 2: The distributions of time gaps in seconds between all movements (A) and explorations (B) for two stages. Both plots are shown in log-log scale. Please note that the two curves mostly overlap in plot (A), indicating little difference between the first and second stages in terms of the distribution of time gaps between all movements. While (B) reveals that the time gaps between explorations are significantly shorter in the first stage.

first activities) up to 2014 Dec[3]. We further filter out inactive users who posted less than 100 activities or joined less than five online communities. We also follow previous work to exclude non-human accounts by detecting those accounts that contain certain terms such as "Bots", "Moderator" and so on, as well as deleted users [18]. In total, our final dataset contains 16 million activities, forming the online trajectories of 57 thousand users. We will specify in the corresponding sections if more constraints are applied to the data.

3 PREFERENTIAL RETURN MODEL

In this section, we introduce our ***preferential return model***. We first model the probability that a user explores a new online community at each move, denoted as P_{new}. Since return and exploration are complementary, the probability of the user returning to a visited community is naturally $1 - P_{new}$. We find evidence that P_{new} decreases over time, and the decreasing tendency can be well modelled with the number of unique communities the user has visited. We then model, given that the user choose to return, the probability that she returns to a specific visited community i, denoted as Π_i.

3.1 Relation between Return and Exploration

According to our experiments, as well as several previous studies [8, 18, 48], the number of unique communities visited by a user, denoted as S, follows

$$S \sim t^{\mu}, \tag{1}$$

where t is the time the user has spent wandering in cyberspace (Figure 1). The parameter μ is found to be smaller than one [18], and empirically estimated to be around 0.4 from our data. The fact that $\mu < 1$ indicates that the exploration of new communities of users shows a decreasing tendency over time. Two possible reasons could cause the tendency: 1) the probability of user exploring new communities, P_{new}, decreases, or 2) P_{new} stays constant, but the user activity levels reduce over time (i.e. performing fewer movements). To find out, we evenly divide user trajectories into two stages according to the number of communities explored. In other words, given that a user has explored S unique communities in total, she explored $S/2$ in each stage, respectively. We examine the time gaps between explorations, as well as the time gaps between all the moves for each stage, respectively. Our experiments show that, in terms of the distributions of the time gaps between all the moves, little difference is observed between two stages (Figure 2 (A)). In other words, statistically, user activity levels are almost unchanged over time. However, the time gaps between explorations in the later stage are significantly longer than in the former stage (Figure 2 (B)). Clearly, these evidences suggest that although users stay active at the same level, the probabilities of exploring new communities decrease over time.

To model the decreasing probability, we associate P_{new} with S. The idea is rooted from our assumption that the more communities a user has visited, the less likely she might want to explore a new one in her next move. Therefore, we formally model P_{new} as

$$P_{new} = \rho S^{-\gamma}. \tag{2}$$

We then quantitatively validate this formula using empirical data. However, it is not practically feasible to estimate P_{new} for every move a user makes. Fortunately, P_{new} can be also viewed as the

[3]We are aware of that our filtering method could still include some users who joined Reddit before 2012 July, but we assume that the amount of such users are few and have little impact on our experiments.

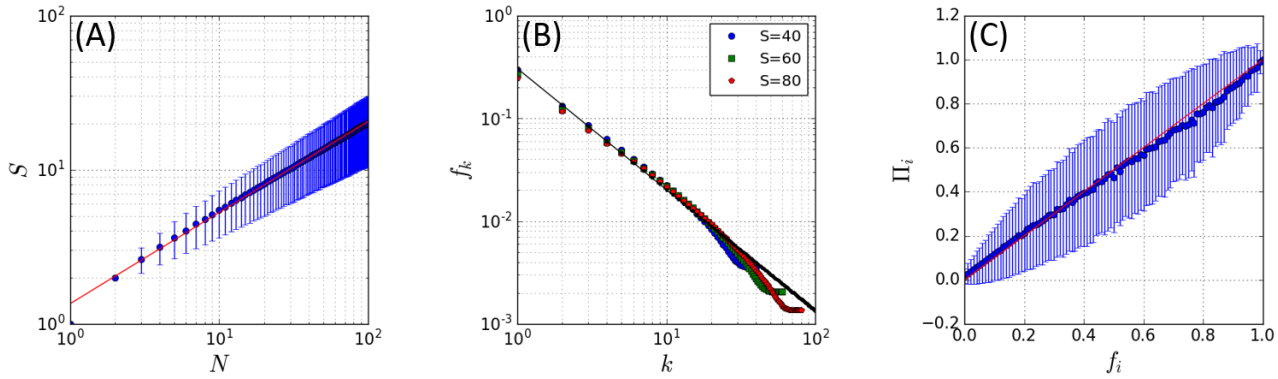

Figure 3: (A) The distribution of the number of uniquely visited communities S over the number of total movements N averaged on all the users, and the error bars indicate the standard deviations. The fitted line in log-log scale indicates that power-law distribution of S over N, as suggested in Formula (4). (B) The relation between visit frequency f_k and visit ranking k for three user groups of different S. The plot shows that Zipf's law holds true, regardless of the number of unique communities a user has visited. (C) The relation between Π_i and f_i averaged on all the user, and the error bars show the standard deviations. The fitted line has a slope of 0.96 and an intercept of 0.005, indicating the equivalence of Formula (6) holds true.

change rate of S over the total number of moves, denoted as N [40][4]. Formally, we can rewrite the equivalence as

$$P_{new} = \frac{\mathrm{d}S}{\mathrm{d}N}. \tag{3}$$

Combining Formulas (3) and (2), we can derive our hypothesized relation between S and N:

$$S = (1+\gamma)(\rho N)^{1/(1+\gamma)}. \tag{4}$$

Since both S and N can be observed from data, we therefore validate our modelling on P_{new} (i.e. Formula 2) by empirically showing that Formula (4) holds true.

Formula (4) suggests a power-law distribution of S over N, i.e. $log(S) \propto log(N)$. We calculate the Pearson Correlation Coefficient between $log(S)$ and $log(N)$ for each user in our dataset. Our experiments report that for more than 96% of the users, the coefficients are larger than 0.9 (p-values < 0.001). In other words, Formula (4) holds true for most users. We further plot this distribution averaged on all the users in Figure 3 (A), and clearly observe a straight line in log-log scale. Therefore, our hypothesis on P_{new} is validated as well. To be specific, we can conclude that, as the number of visited communities increases, the probability of an individual exploring new communities decreases exponentially. Practically, suggested by Formula (4), we are able to determine γ and ρ for each user from the parameters of the straight line in log-log scale[5], and study human mobility in cyberspace at an individual granularity (Section 6).

3.2 Probability of Returning to a Community

Modelling P_{new} allows us to predict how likely an individual would explore a new community, or return to a visited community. We next study the probability of an individual returning to a specific visited community i, i.e. Π_i. It is reported that the frequency f_k of a user's kth most visited community follows Zipf's law [18, 54], and formally,

$$f_k \sim k^{-\zeta}. \tag{5}$$

Figure 3 (B) shows that the this law also appears in our data. The Zipf's law indicates users' imbalanced probabilities of returning to visited communities. Such a "the rich get richer" observation inspires us to model Π_i as the proportion of visits paid to this community previously, denoted as f_i. Formally,

$$\Pi_i = f_i. \tag{6}$$

Theoretically, given P_{new} being modelled as $\rho S^{-\gamma}$, this hypothesis on Π_i can lead to the Zipf's law property in Formula (5), providing evidence for the rationale of Formula (6)[6]. Empirically, We further validate this formula using our data. Since it is infeasible to directly estimate Π_i for each move, we make an approximation in our experiments. To be specific, we cut user trajectories into slices of length L. Within each slice, we estimate Π_i as the percentage of visits to community i. Then within the previous slice, we compute f_i, the frequency to the same community. We set L as 100 for a robust estimate for Π_i[7], and conduct our experiments on totally 23 thousand users whose trajectories cover at least two slices (i.e. total movements N large than 200). We plot the relation between Π_i and f_i in Figure 3 (C). From the plot, we can observe that the relation is linear with a slope very close to 1, indicating that Formula (6) holds true. An interesting observation is the spindle-shaped errors of the fitted straight line – the equivalence between Π_i and f_i is more stable when f_i approaches either end (1 or 0). This is because

[4]Generally speaking, this is because the probability of visiting a new community at each move is equal to the "amount of change" of S brought by each move. For more information, please refer to the supplementary material of [40]

[5]Simply, the slope of the straight line is $1/(1+\gamma)$, and the intercept is $log(\rho)/(1+\gamma) - log(1+\gamma)$.

[6]A detailed proof is beyond the scope of this paper. Please refer to [40] for more information.

[7]Naturally, larger values of L lead to more precise estimates of Π_i. However, since Π_i and f_i are estimated within two different slices, larger L also introduces more temporal differences to the slices, and therefore, introduces larger errors to our validation. We experiment with L from a reasonable range of 50 to 200, and obtain similar results.

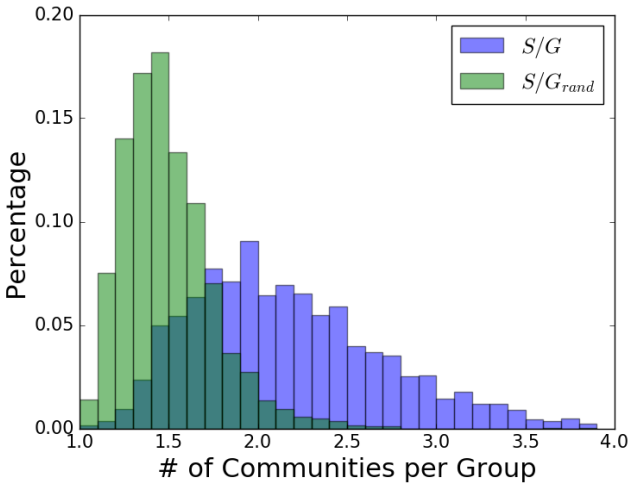

Figure 4: Histogram of the average numbers of online communities a user explored per community-group under our group assignments (S/G) versus random group assignments (S/G_{rand}).

our approximation might introduce some amounts of noises into the fitting, unless the user previously rarely visits a community ($f_i \to 0$) or visits the community very frequently ($f_i \to 1$).

4 PREFERENTIAL EXPLORATION MODEL

The preferential return model characterizes the statistic patterns of ***return*** movement in cyberspace. However, this model does not describe if a user chooses to explore, which new community she will visit. In this section, we introduce our *preferential exploration model* for characterizing ***exploration*** movement. The model is inspired by the fact that the communities a user visit usually can be categorized into several community-groups [12, 53]. Table 1 shows the exploration traces of three randomly selected users from our data. For example, the thrird user in the table explored $S = 28$ unique online communities, while most of these communities are mainly from five community-groups. To be specific, community *sushi*, *asianeats* (asian eats), *tipofmytongue* (tip of my tongue), and *ramen* are clearly related to ***Food***. Similarly, *Boxing*, *kungfu*, *loseit* (lose it), and so on are all communities on ***Exercises***. Such a grouping property suggests that the explorations of a user are not independent – previously explored community-groups could provide further information for later explorations. Following this idea, we investigate exploration movement from a community-group perspective. In particular, we first categorize all the online communities into community-groups. Then we study how users select unvisited communities from community-groups at each exploration.

4.1 Extracting Community-groups

In this work, we adopt a dendrogram-based method to extract community-groups, as suggested in many community detection studies [22, 38]. We apply Pointwise Mutual Information (PMI) as the similarity metric for clustering [42]. To be specific, for two online community C_i and C_j, let U_{C_i} and U_{C_j} denote their user sets, respectively. We compute the similarity between C_i and C_j, as $pmi(U_{C_i}, U_{C_j})$, and formally,

$$pmi(U_{C_i}, U_{C_j}) = log\frac{|U_{C_i} \cap U_{C_j}|}{|U_{C_i} \cup U_{C_j}|}, \tag{7}$$

where $|\cdot|$ denotes the size of a set. PMI computes the similarity between communities from a user co-occurrence perspective. Compared with language-based metrics that are usually adopted for measuring online community similarity [43, 45], PMI better quantifies the likelihood of users moving across communities. For example, *nyc* and *boston* are two typical city-specific communities on Reddit. Since the visitors to city-specific communities are mainly local residents [18], typical users of *nyc* or *boston* are unlikely to pay visit to the other community. Therefore, the similarity between *nyc* and *boston* should be relatively low from a movement perspective, as indicated by a low value of $pmi(nyc, boston)$. On the contrary, according to our experiments, language-based metrics, such as similarities between doc2vec embedding [23], suggest high similarity between the two communities, because of their similarly city-related topics. We set the number of community-groups as 300 for clustering[8], and extract flat group structures for online communities, i.e., each online community is assigned to only one community-group. Table 2 shows some examples of the extracted community-groups. We define that a community-group is explored by a user if the user explores one or more online communities of the group.

We further examine our community-group assignments by comparing the results with random group assignments. Specifically, we randomly cluster all the communities into the same number of groups, and the distribution of group sizes (i.e. number of online communities in each group) is also set to the same as our group assignments. Let G denote the number of community-groups a user explored under our assignments and G_{rand} denote the number of random community-groups a user explored. If our assignments capture the grouping property of user explorations, i.e., user explorations concentrate on several community-groups, then the average number of online communities a user visited per community-group, i.e. S/G, should be significantly larger than S/G_{rand}. We compute both S/G and S/G_{rand} for each user, and conduct a t-test to compare the two values. The test result indicates that a user explores significantly more online communities per community-group under our assignments than random group assignments with a p-value < 0.001. Therefore, our community-group assignments are validated (see Figure 4 for more details).

4.2 Exploring New Community-groups

With the group labels being assigned to online communities, we study the probability of a user exploring a community in a community-group that she has never explored before. We denote this probability as Q_{new}, and the probability of exploring a community in her explored community-groups is thus $1 - Q_{new}$. Similar to our inference on P_{new}, the probability Q_{new} equals to the change rate of the number of unique community-groups a user explores, G, over the number of unique communities she visits, S. Formally, the

[8]Using different settings of community-group numbers leads to same statistic patterns of exploration movement, but with slightly different parameters.

Table 1: Exploration traces of three randomly selected users from our dataset. We also indicate the community-groups each user explored. The community-groups are automatically extracted from all the online communities using a clustering method, and the titles are manually labeled. We use different colors and superscripts in the exploration traces to label which community-group an online community belongs to. The online communities that do not share group membership with any other communities in a trace are colored in black, and are not denoted with superscripts. Also, please note that *AUS.*, *RELP.*, and *EXER.* are short for Australia, Relationships, and Exercises, respectively.

User	*Exploration Traces of Online Communities*	*Comm-groups*
1	*espnyankees*[1] → *NYYankees*[1] → *Libertarian*[2] → *nyjets*[1] → *rangers*[1] → *funny* → *baseball*[1] → *nfl*[1] → *politics*[2] → *nba*[1] → *ShitPoliticsSays*[2] → *NYKnicks*[1].	1. ***Sports*** 2. ***Politics***
2	*subaru*[1] → *Android*[2] → *Autos*[1] → *S2000*[1] → *sysadmin* → *CarAV*[1] → *DIY*[3] → *Documentaries* → *BuyItForLife*[3] → *Ford*[1] → *commandline*[2] → *thewholecar*[1] → *cars*[1] → *UserCars*[1] → *fixit*[3] → *Nexus5*[2] → *programming*[2] → *Shitty_Car_Mods*[1] → *HomeNetworking*[3] → *DesignMyRoom*[3].	1. ***Cars*** 2. ***Tech*** 3. ***DIY***
3	*sushi*[1] → *melbourne*[2] → *todayilearned* → *AustralianPolitics*[2] → *books* → *relationship_advice*[3] → *Boxing*[4] → *amateur_boxing*[4] → *Tobacco*[5] → *ForeverAlone*[3] → *asianeats*[1] → *bodybuilding*[4] → *tipofmytongue*[1] → *kungfu*[4] → *homegym*[4] → *wallpaper* → *ForeverAloneDating*[3] → *Cigarettes*[5] → *MeetPeople*[3] → *loseit*[4] → *CigarMarket*[5] → *dating*[3] → *circlejerkaustralia*[2] → *MakeNewFriendsHere*[3] → *Zippo*[5] → *ramen*[1] → *weightlifting*[4] → *MMA*[4].	1. ***Food*** 2. ***Aus.*** 3. ***RELP.*** 4. ***EXER.*** 5. ***Smoke***

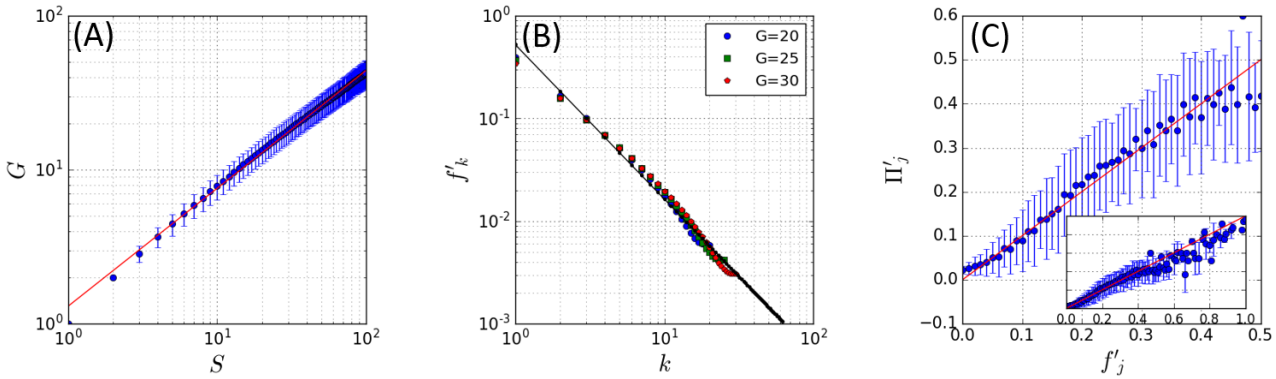

Figure 5: (A) The distribution of the number of unique explored community-groups G over the number of explored communities S averaged over all the users. The plot is in log-log scale, and the error bars indicate the standard deviations. (B) Zipf's law between the exploration frequency to community-groups, f'_k, and the exploration ranking k. (C) The relation between Π'_j and f'_j. The main plot and the inset plot shows the ranges of [0, 0.5] and [0, 1] for f'_j, respectively. Both plots are averaged over all the users, and the error bars show the standard deviations.

equivalence is

$$Q_{new} = \frac{dG}{dS}. \tag{8}$$

We therefore investigate the relation between G and S from user trajectories. Very interestingly, we also discover a power-law distribution of G over S, i.e. $log(G) \propto log(S)$. Specifically, the Pearson Correlation Coefficients between $log(G)$ and $log(S)$ of over 91% users are higher than 0.9, with p-values smaller than 0.001. We plot this distribution in log-log scale in Figure 5 (A).

Given the power-law distribution, as well as Formula (8), we are able to infer the expression of Q_{new}, and formally,

$$Q_{new} = \rho' G^{-\gamma'}. \tag{9}$$

In other words, the probability of exploring a new community-group exponentially decreases with the number of unique groups previously explored. This finding makes interesting but obvious sense – a user's exploration in cyberspace gradually converges to the several categories (i.e., community-groups) she is interested in, and consequently, the more community-groups she has explored, the less likely she will explore a new one. Furthermore, the power-law distribution of G over S can be simply parameterized as

$$G = (1 + \gamma')(\rho' S)^{1/(1+\gamma')}. \tag{10}$$

Please note that, the relation between S and G can be empirically estimated from user trajectories. Therefore, Formula (10) allows us to compute both ρ' and γ' for each user, and to quantify individual exploration patterns (Section 6).

Table 2: Five extracted community-groups using an automatic clustering method with manually labeled titles. The first three groups –*Food*, *Tech*, and *Cars* in the table, are already mentioned in Table 1, and we show more online communities from these community-groups here. We also list two new community-groups – *Wine* and *Asia* in the table.

Title	*Online Communities*
Tech	*Fedora, storage, SQLServer, webdesign, PowerShell, exchangeserver, ipv6, gnu, aws, hacking, git, Cisco, cyberlaws, redhat, Database ,django, docker,* etc.
Cars	*Porsche, BMW, racing, 350z, 4x4, spotted, formula1, RX7, infiniti, LandRover, ATV, Drifting, driving, TopGear, , RatRod, Jeep, HotWheels, carporn, E30.* etc.
Food	*JapaneseFood, grilling, IndianFood, Pizza, recipes steak, BBQ, KoreanFood, streeteats, Cheese, HotPeppers, grilledcheese, slowcooking, meat, hotsauce, Bacon,* etc.
Wine	*tequila, Gin, alcohol, cocktails, liquor, rum, Scotch, vodka, Whiskyporn, beer, wine, bourbon, drunk, worldwhisky, brewing, whisky, beerwithaview,* etc.
Asia	*Korean, China, Osaka, japannews, VietNam, Bangkok, japanesemusic, Tokyo, seoul, JapanTravel, Chinese, taiwan, beijing, Chinavisa, shanghai, cambodia,* etc.

4.3 Exploring Previous Community-groups

Next, we investigate how users select new communities to explore from the community-groups explored before. We start with the exploration frequency a user exercises in each community-group, denoted as f'. For the first time, we uncover that Zipf's law also applies to f'. Generally speaking, users perform exponentially more explorations in their preferred community-groups, resulting in very imbalanced exploration frequencies among community-groups. Formally, the exploration frequency f'_k of the kth most explored community-group satisfies

$$f'_k \sim k^{-\zeta'}. \tag{11}$$

Figure 5 (B) shows the Zipf's law for the users of different numbers of explored community-groups. The exponentially uneven exploration frequency indicates that "the rich get richer" is also reflected in user exploration at the community-group level. Therefore, we naturally model the probability a user explores a community in her explored community-group j, denoted as Π'_j, as the exploration frequency of this community-group f'_j, and formally,

$$\Pi'_j = f'_j. \tag{12}$$

We empirically validate $\Pi'_j = f'_j$ by employing the same method for validating $\Pi_i = f_i$. Specifically, we divide user exploration trajectories into slices of length L'. Within a slice, we estimate Π'_j as the percentage of explorations to community-group j, and within the previous slice we compute f'_j for the same community-group. Although a large L' makes the estimations more precise for users who have sufficiently long exploration trajectories, it also significantly reduces the total number of users for our experiments. Therefore, we make a trade-off, and set $L' = 20$. We then filter out the users who explore fewer than 40 communities ($S < 40$), and retain 4 thousand users. We plot our results in Figure 5 (C). Despite the relatively large errors caused by the value of L', we still see the approximate equivalence between Π'_j and f'_j, which validates Formula (12). Note that the errors increase as f'_j increases, especially for $f'_j > 0.5$ (see the inset of Figure 5 (C)). This is due to the fact that a user can explore at most one community-group with a frequency larger than 0.5. Meanwhile, the total number of users is relatively small. Hence the errors for $f'_j > 0.5$ are relatively large.

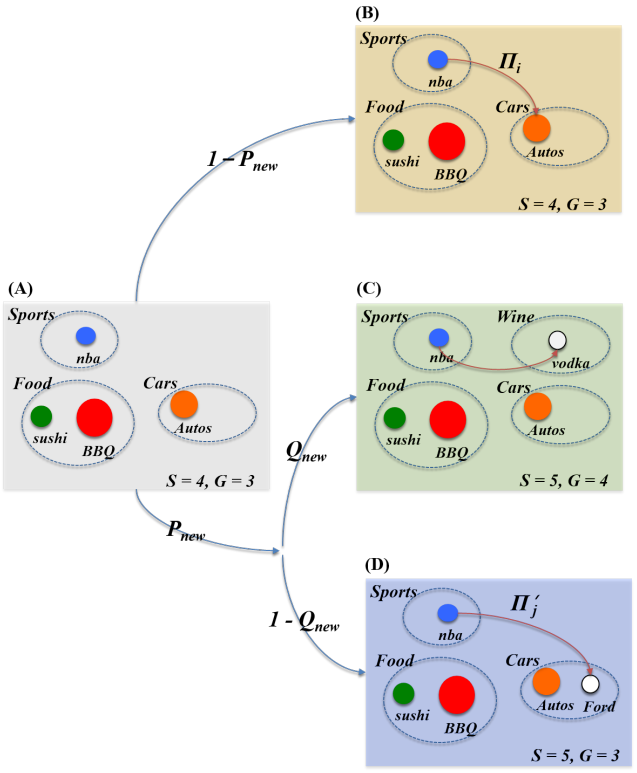

Figure 6: An example of an imaginary user's movements across online communities, illustrating our overall model. From the initial stage (A), the user first chooses between returning to a previously visited community (B) or exploring a new one. If the user chooses to explore, she further decides between exploring a new community-group (C) or selecting a new community from a previously explored community-group (D).

5 OVERALL MODEL

Combining both ***preferential return model*** and ***preferential exploration model***, we are able to cover two basic types of movements across online communities, and obtain an overall model for human mobility in cyberspace.

5.1 An illustrative example of our model

We summarize our overall model using an example of an imaginary user moving across online communities (Figure 6):

- At the initial stage (Figure 6 (A)), the user has visited four communities, belonging to three communities groups ($S = 4$ and $G = 3$): *nba* – ***Sports*** group, *sushi* and *BBQ* –***Food*** group, and *Autos* – ***Cars*** group.
- The user first chooses to perform either exploration or return, with probability P_{new} and $1 - P_{new}$, respectively. Our model indicates that probability P_{new} can be modelled as $\rho S^{-\gamma}$.
- If she chooses to return, she returns to community i with probability Π_i, and the probability can be modeled as the visit frequency to this community f_i. Figure 6 (B) illustrates that the user returns to her previously visited community – *Autos*. Both S and G are unchanged in this case.
- Otherwise, she further chooses to explore an unvisited community in a new or an explored community-group, with probability Q_{new} and $1 - Q_{new}$, respectively. We model Q_{new} as $\rho' G^{-\gamma'}$ in our model. Figure 6 (C) shows that the user explores community *vodka* in a new community-group ***Wine***. In this case, both S and G are increased by one.
- Furthermore, the probability of the user exploring a new community in a previously explored community-group j, denoted as Π'_j, is proportional to the exploration frequency of this community-group f'_j. Figure 6 (D) shows that the user explores a new community *Ford* in her explored community-group ***Cars***. Therefore, S increases to 5, and G remains 3.

5.2 The Preferential Principle

This overall model uncovers two interesting resemblances with respect to human dynamics: 1) the resemblance between human movements in cyber and physical spaces, and 2) the resemblance between online return and exploration movements. First, according to the physical preferential model [30], the relation between return and exploration in physical world shares the same form of Formula (2). Meanwhile, the probability of returning to a physical location is also proportional to the previous visit frequency of this location (Formula 6). In other words, although the two spaces are fundamentally different, people follow a similar "preferential principle" in both spaces – the probability of visiting new communities/locations decreases over time, and people pay more and more visits to the communities/locations they are fond of. Second, it is worth noting the similar mathematical forms of our preferential return and exploration models. The similarities further suggest that the "preferential principle" not only governs how people return to visited online communities, but also applies to the explorations to new communities. To be specific, people explore new community guided by their preference to community-groups – they gradually lose interests in exploring new community-groups, and perform more and more explorations to their preferred groups.

Such a "preferential principle" in both cyber and physical spaces, and for both return and exploration movements, leads to the frequently observed Matthew effect [26] ("the rich get richer" phenomenon) in human mobility patterns. Given the fact that Matthew effect is also observed in many other areas, such as economy and sociology [33, 41], it is worth questioning that if this principle is one of the fundamental rules of human dynamics, and applicable to other human activities. We will save these very interesting extensions for our future work.

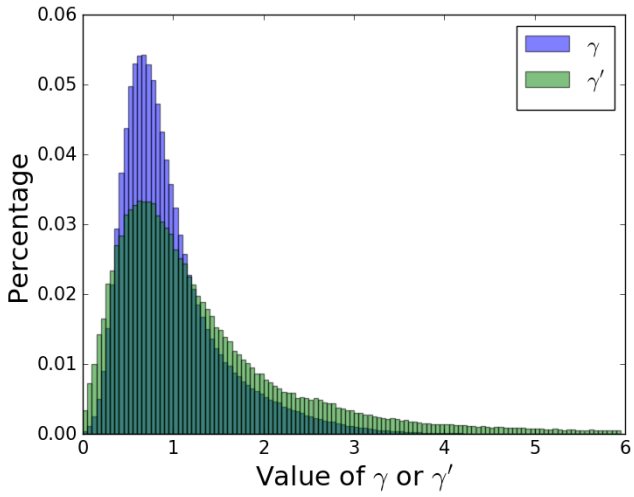

Figure 7: Histogram of γ versus γ' for all the users .

6 HUMAN DYNAMICS IN CYBERSPACE

Our models reveal that user intentions to explore new online communities and community-groups both decrease over time. From Formulas (4) and (10), we can separately derive

$$log(S) = 1/(1+\gamma)log(N) + const, \tag{13}$$

$$log(G) = 1/(1+\gamma')log(S) + const. \tag{14}$$

Apparently, parameter γ quantifies how strong users' intentions are to explore new communities. A larger γ indicates a weaker intention – the increase of number of unique communities S is slower as the total movements N increases. Similarly, parameter γ' quantifies how strong users' intentions are to explore new community-groups. Therefore, in this section, we focus on γ and γ', and study human dynamics in cyberspace.

6.1 Intentions to Explore

We compute both γ and γ' for all the users in our dataset. Our results report that the mean and median of γ are 0.97 and 0.82, respectively; and the mean and median of γ' are 2.52 and 1.17, respectively. We then conduct a t-test between γ and γ', and the test result indicates that γ' is significantly larger than γ (t-statistic = 41.9, and p-value < 0.001). Figure 7 shows the distributions of both parameters. Recall that, in physical space, the relation between return and exploration are the same as in cyberspace. According to the previous research [40], the intention to explore new places in physical space, parameterized by γ_p, is estimated around 0.2. Therefore, we have

$$\gamma_p < \gamma < \gamma' \tag{15}$$

Formula (15) reveals previously unknown insights of human intention to explore. On one hand, from $\gamma_p < \gamma$, we learn that people are more likely to explore in physical space than in cyberspace. One possible reason for the difference lies in the evolution of physical

Table 3: Top five positively and negatively correlated online communities to γ. We list the communities along with their user sizes and the values of the coefficients.

Subreddit	*User Size*	*Coefficient*
Top five positively correlated online communities		
Albuquerque	11,970	0.49
letsgofish	3,211	0.39
tulsa	10,324	0.37
deephouse	23,674	0.33
TankPorn	47,351	0.33
Top five negatively correlated online communities		
TrueAskReddit	158,669	-0.38
poker	84,276	-0.34
baseball	822,279	-0.34
Needafriend	68,294	-0.31
AskHistorians	836,844	-0.29

locations and online communities. Online communities are self-updating – people communicate and interact in communities, and keep posting new contents [24]. On the contrary, physical locations are usually associated with fixed functionality, and change much slower over time. Therefore, people can still be exposed to new contents even if they return frequently to a visited communities in cyberspace, but they might feel bored if they do not explore new places in physical space. On the other hand, the intention to explore new community-groups is the lowest among three, suggested by the large value of γ'. This result is consistent with the psychology theories that describe human cognitive upper limit of handling interests [27, 28]. In other words, when people have explored sufficient community-groups to fulfill their interests, they mainly focus on exploring the new communities within the explored groups instead of opening up new community-groups.

In a nutshell, human dynamics in cyberspace shares many similar principles with physical space while maintaining several unique characteristics.

6.2 Relations between Intentions and Contents

With the knowledge of user intentions to explore, a natural follow-up question is if the intentions are related to the communities or community-groups a user visits. For example, we wonder if people with lower intentions visit different communities from people with higher intentions. To answer the questions, we conduct two linear regression analyses. The dependent valuables of the analyses are γ and γ' of users, and the independent valuables are the online communities and community-groups the users visit for the two analyses, respectively. The two analyses report quite different results. The first analysis reports a high r^2 of over 0.67. This suggests that more than 67% of the variance of γ can be explained by the online communities a user visits. In contrast, the second analysis reports a quite low r^2 of 0.11, indicating that the intention of exploring new community-groups can be barely explained by users' visit histories of community-groups. The analysis results probably suggest human's different cognitive views of single online community and community-groups. We leave the investigation of such differences for the future work.

Furthermore, from the results of the linear regression analysis of γ, we are able to learn the communities that are negatively and positively correlated with γ. Specifically, for each independent valuable (i.e., online community), the linear regression calculates a coefficient and a p-value. The coefficient indicates how correlated an independent valuable is to the dependent valuable, and the p-value suggests how significant the correlation is. We first filter out all the online communities whose p-values are larger than 0.05, and obtain 381 communities positively correlated with γ, and another 427 communities negatively correlated. We list the top five most positively and negatively correlated communities in Table 3. Please note that, positive correlations to γ imply that people who visit these communities prefer to return, and oppositely, negative correlations imply that their visitors prefer to explore.

We can observe that positively and negatively correlated communities are quite different in several aspects. First, positive correlated communities are usually of smaller sizes. This difference can be directly read from the table, and is also confirmed by a t-test (t-statistic = 3.8, p-value < 0.001). Second, the two types of communities are of quite divergent topics. The positively correlated communities usually focus on specific and niche topics. For example, the first (*Albuquerque*) and the third (*tulsa*) are local communities of two American cities[9]. In fact, many other city-specific communities also appear to be positively correlated. The topics of the other three positively correlated communities are also quite specific. For example, *TankPorn* is a community for sharing pictures of tanks, and *letsgofish* focuses on Miami Marlins – an American professional baseball team based in Miami, Florida. While the topics of negatively correlated communities are broader and more interactive. For example, *TrueAskReddit* and *AskHistorians* are both Q&A communities. Meanwhile, *Needafriend* is a community for building up friendships. The contents in these communities involve much more interactions between users. The remaining two communities are both related to general topics – discussions on *poker* and *baseball*.

7 DISCUSSION & FUTURE WORK

Different from previous studies on human online activities, in this work we employ a novel mobility perspective, and decompose online activities into return and exploration movements. Our models indicate that both online return and exploration movements follow a preferential principle. Such a finding further suggests several valuable research questions. Since return movement in physical space reportedly also exhibits the same properties, a question worth studying is if humans explore physical space in a similar preferential way. Furthermore, we also wonder if the principle could extend to other types of human activities. For example, we can study if certain human daily activities, e.g., listening to music, can also be modelled the same way. To be specific, it is possible that people re-listen to their preferred music many times (similar to return movements); and the more music they have listened, the less likely they will seek for new music (similar to exploration movements).

[9]The two cities are Albuquerque, New Mexico, and Tulsa, Oklahoma, respectively.

Moreover, given the fact that the preferential returns are never observed in animal mobility [5], we are also curious if such a preferential principle is one of the fundamental, and meanwhile, unique principles of human dynamics.

Our findings in this paper also imply rich practical applications. For example, we uncover that human explorations in cyberspace are guided by an underlying group structure and related to previously explored community-groups. Meanwhile, our models quantify individual intentions to explore both online communities (γ) and community-groups (γ'). This knowledge could be applied to predicting users' next exploration, and implementing a more sophisticated recommendation system. For example, it is possible to deploy more personalized recommendations based on user exploration intentions. Specifically, for users with a low γ but a high γ', the system can focus on recommending communities within their explored groups; and for users with a low γ and also a low γ', the system may recommend more communities from new community-groups.

Furthermore, our preferential exploration model is built on a flat community-group structure in cyberspace. We extract the group structure from data using an automatic clustering method. Our experiments report that different settings of the number of groups result in slightly different parameters (i.e., γ' and ρ'), but the statistical patterns of exploration in cyberspace are consistent. It is worth noting that many previous studies suggest that the group structure revealed in human activities are in fact hierarchical and overlapping [30, 38]. Flat community-group structures with different group numbers can be regarded as rooted from the same general hierarchy but with different cutting thresholds, and are special cases of overlapping group structures. These facts suggest a more ubiquitous possibility of modelling human online exploration – humans explore communities following an implicit "road-map" of cyberspace. To be specific, in this new model, a global hierarchy defines the positional relations and the cost of moving across online communities. In other words, the hard boundaries of community-groups are replaced by continuous "distances" in cyberspace. We can then naturally associate the probability of exploration with a distance measure and obtain a more generic model without manually setting any thresholds.

8 RELATED WORK

Since Rotman et al. first suggest that people interact and share purpose in online groups, and therefore form sociological communities [37], there has been much work studying human activities across online communities. It is common that people join multiple communities on online platforms [6], and user participation in multiple communities reportedly benefits the survival of the communities [55]. Danescu-Niculescu-Mizil et al. study the interaction between users and online communities, and discover that there are different stages of users joining online communities from a linguistic perspective [10]. Following a similar approach, Tan et al. investigate the language change of users while joining multiple communities, and report that people tend to visit less and less popular communities [44]. Furthermore, how communities affect user actives are also studied. For example, Hamilton et al. study the user loyalty to multiple communities, and reveal the relation between user engagement and community characteristics [15]. Kumar et al. report that communities interact, and even conflict with each others, and most conflicts are started by a small number of specific online communities [21].

Our work is also closely related to the work studying human mobility. Many models have been developed in physics to investigate the mechanisms of the movements across locations in real word [14, 46, 49]. For example, Barabasi et al. propose a priority queue mechanism to model [1], and Ravasz et al. suggest a hierarchical organization [34] to explain the observed burstiness property of human mobility. Song et al. first propose the preferential return model for quantifying the return and exploration movements in physical world [40]. Such a model is then extended by much follow-up work [3, 25]. For example, Pappalardo et al. develop a "returner and explorer dichotomy" for differentiating people of divergent mobility patterns [32].

The similarities between human activities across online communities and movements in physical space have been discussed in some recent work [52, 54]. For example, Hu et al. introduce an analogy between physical space and cyberspace, and leverage framework designed from movements across locations in real world to study mobility patterns across online communities [18]. However, these studies mostly focus on statistical resemblances, instead of investigating the underlying rules that lead to the similarities [2]. Different from the previous work, in this work we empirically show that the mechanisms of movements in both spaces also share interesting similarities.

9 CONCLUSIONS

Given the increasing involvement and significance of cyberspace in people's daily lives, it is important to model the underlying patterns and principles of human online activities. In this work, we have systematically studied human activities across online communities from a mobility perspective. We first propose the preferential return model, which characterizes how people choose between returning to a previously visited community or exploring a new one. Then the model quantifies, if the user decides to return, how likely she is going to return to a specific community. With the return movements being modelled, we further propose the preferential exploration model. This model analyzes human online exploration movements from a novel community-group perspective, and quantitatively describes the processes of people selecting unvisited communities from community-groups. The two models are substantially validated by statistically solid evidences on a large real world data set. Combining these two models, we are able to reveal an interesting preferential principle of human cyberspace dynamics. Generally speaking, the principle suggests that people gradually lose interest in exploring new physical or cyber locations, and more and more often return to the location they are fond of. Such a principle implies a rich variety of future work (as stated in Section 7).

We have further applied our models to analyzing human online trajectories at the individual level, and reveal previously undiscovered insights about human mobility in cyberspace. We quantify user intention to explore in cyberspace and find that, interestingly, people are more likely to explore in physical world. Moreover, we investigate the relations between online communities and user intention to explore. Our experiments report that people who prefer

to return tend to visit communities of smaller size and specific topics, while people who prefer to explore tend to visit communities of larger size and interactive topics.

REFERENCES

[1] Albert-Laszlo Barabasi. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435, 7039 (2005), 207.

[2] Hugo S Barbosa, Fernando B de Lima Neto, Alexandre Evsukoff, and Ronaldo Menezes. 2016. Returners and Explorers Dichotomy in Web Browsing Behavior: A Human Mobility Approach. In *Complex Networks VII*. Springer, 173–184.

[3] Marc Barthélemy. 2011. Spatial networks. *Physics Reports* 499, 1-3 (2011), 1–101.

[4] Václav Belák, Samantha Lam, and Conor Hayes. 2012. Cross-Community Influence in Discussion Fora. *ICWSM* 12 (2012), 34–41.

[5] Clifford T Brown, Larry S Liebovitch, and Rachel Glendon. 2007. Lévy flights in Dobe Ju/âĂŹhoansi foraging patterns. *Human Ecology* 35, 1 (2007), 129–138.

[6] Jeffrey Chan, Conor Hayes, and Elizabeth M Daly. 2010. Decomposing Discussion Forums and Boards Using User Roles. *ICWSM* 10 (2010), 215–218.

[7] Anna Chmiel, Kamila Kowalska, and Janusz A Hołyst. 2009. Scaling of human behavior during portal browsing. *Physical Review E* 80, 6 (2009), 066122.

[8] Anna Chmiel, Julian Sienkiewicz, Mike Thelwall, Georgios Paltoglou, Kevan Buckley, Arvid Kappas, and Janusz A Hołyst. 2011. Collective emotions online and their influence on community life. *PloS one* 6, 7 (2011), e22207.

[9] Eunjoon Cho, Seth A Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1082–1090.

[10] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 307–318.

[11] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3 (2013), 1376.

[12] Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. 2006. Topical interests and the mitigation of search engine bias. *Proceedings of the National Academy of Sciences* 103, 34 (2006), 12684–12689.

[13] Michelle Girvan and Mark EJ Newman. 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences* 99, 12 (2002), 7821–7826.

[14] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *nature* 453, 7196 (2008), 779.

[15] William L Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in online communities. In *Proceedings of the... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media*, Vol. 2017. NIH Public Access, 540.

[16] Lei Hou, Xue Pan, Qiang Guo, and Jian-Guo Liu. 2014. Memory effect of the online user preference. *Scientific reports* 4 (2014), 6560.

[17] Tianran Hu, Eric Bigelow, Jiebo Luo, and Henry Kautz. 2017. Tales of two cities: Using social media to understand idiosyncratic lifestyles in distinctive metropolitan areas. *IEEE Transactions on Big Data* 3, 1 (2017), 55–66.

[18] Tianran Hu, Jiebo Luo, and Wei Liu. 2018. Life in the" Matrix": Human Mobility Patterns in the Cyber Space. *ICWSM* (2018).

[19] Sanjay Ram Kairam, Dan J Wang, and Jure Leskovec. 2012. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 673–682.

[20] Robert E Kraut and Paul Resnick. 2012. *Building successful online communities: Evidence-based social design*. Mit Press.

[21] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 933–943.

[22] Peter Langfelder, Bin Zhang, and Steve Horvath. 2007. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24, 5 (2007), 719–720.

[23] Jey Han Lau and Timothy Baldwin. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. 78–86.

[24] Zhiyuan Lin, Niloufar Salehi, Bowen Yao, Yiqi Chen, and Michael S Bernstein. 2017. Better When It Was Smaller? Community Content and Behavior After Massive Growth.. In *ICWSM*. 132–141.

[25] Xin Lu, Linus Bengtsson, and Petter Holme. 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences* 109, 29 (2012), 11576–11581.

[26] Robert K Merton. 1968. The Matthew effect in science: The reward and communication systems of science are considered. *Science* 159, 3810 (1968), 56–63.

[27] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

[28] Ulric Neisser. 2014. *Cognitive psychology: Classic edition*. Psychology Press.

[29] Mark EJ Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E* 69, 2 (2004), 026113.

[30] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 7043 (2005), 814.

[31] Pietro Panzarasa, Tore Opsahl, and Kathleen M Carley. 2009. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology* 60, 5 (2009), 911–932.

[32] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. 2015. Returners and explorers dichotomy in human mobility. *Nature communications* 6 (2015), 8166.

[33] Alexander M Petersen, Woo-Sung Jung, Jae-Suk Yang, and H Eugene Stanley. 2011. Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences* 108, 1 (2011), 18–23.

[34] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. 2002. Hierarchical organization of modularity in metabolic networks. *science* 297, 5586 (2002), 1551–1555.

[35] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World Wide Web*. ACM, 695–704.

[36] Martin Rosvall and Carl T Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105, 4 (2008), 1118–1123.

[37] Dana Rotman, Jennifer Golbeck, and Jennifer Preece. 2009. The community is where the rapport is–on sense and structure in the youtube community. In *Proceedings of the fourth international conference on Communities and technologies*. ACM, 41–50.

[38] Huawei Shen, Xueqi Cheng, Kai Cai, and Mao-Bin Hu. 2009. Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications* 388, 8 (2009), 1706–1712.

[39] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. 2012. A universal model for mobility and migration patterns. *Nature* 484, 7392 (2012), 96.

[40] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. 2010. Modelling the scaling properties of human mobility. *Nature Physics* 6, 10 (2010), 818.

[41] Alan T Sorensen. 2007. Bestseller lists and product variety. *The journal of industrial economics* 55, 4 (2007), 715–738.

[42] Qi Su, Kun Xiang, Houfeng Wang, Bin Sun, and Shiwen Yu. 2006. Using pointwise mutual information to identify implicit features in customer reviews. In *International Conference on Computer Processing of Oriental Languages*. Springer, 22–30.

[43] Chenhao Tan. 2018. Tracing Community Genealogy: How New Communities Emerge from the Old. *arXiv preprint arXiv:1804.01990* (2018).

[44] Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1056–1066.

[45] Trang Tran and Mari Ostendorf. 2016. Characterizing the language of online communities and its relation to community reception. *arXiv preprint arXiv:1609.04779* (2016).

[46] Alexei Vázquez, Joao Gama Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor, and Albert-László Barabási. 2006. Modeling bursts and heavy tails in human dynamics. *Physical Review E* 73, 3 (2006), 036127.

[47] Alex Wang, William L Hamilton, and Jure Leskovec. 2016. Learning linguistic descriptors of user roles in online communities. In *Proceedings of the First Workshop on NLP and Computational Social Science*. 76–85.

[48] Chunyan Wang, Mao Ye, and Bernardo A Huberman. 2012. From user comments to on-line conversations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 244–252.

[49] Ye Wu, Changsong Zhou, Jinghua Xiao, Jürgen Kurths, and Hans Joachim Schellnhuber. 2010. Evidence for a bimodal distribution in human communication. *Proceedings of the national academy of sciences* (2010).

[50] Amy X Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Proceedings of the Eleventh International Conference on Web and Social Media. AAAI Press*.

[51] Justine Zhang, William L Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In *Proceedings of the... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media*, Vol. 2017. NIH Public Access, 377.

[52] Zhi-Dan Zhao, Zi-Gang Huang, Liang Huang, Huan Liu, and Ying-Cheng Lai. 2014. Scaling and correlation of human movements in cyberspace and physical

space. *Physical Review E* 90, 5 (2014), 050802.
[53] Zhi-Dan Zhao, Zimo Yang, Zike Zhang, Tao Zhou, Zi-Gang Huang, and Ying-Cheng Lai. 2013. Emergence of scaling in human-interest dynamics. *Scientific Reports* 3 (2013), 3472.
[54] Zhi-Dan Zhao and Tao Zhou. 2012. Empirical analysis of online human dynamics. *Physica A: Statistical Mechanics and its Applications* 391, 11 (2012), 3308–3315.
[55] Haiyi Zhu, Robert E Kraut, and Aniket Kittur. 2014. The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 281–290.