Demystifying Levy Walk Patterns in Human Walks

Kyunghan Lee†, Seongik Hong††, Seong Joon Kim††,
Injong Rhee†† and Song Chong†

† School of EECS, KAIST, Daejeon, Korea, khlee@netsys.kaist.ac.kr, song@ee.kaist.ac.kr
†† Dept. of Computer Science, NC State University, Raleigh, NC. USA, {shong, sjkim2, rhee}@ncsu.edu

ABSTRACT

This paper reports that bursty hot spot sizes are a key factor in causing the heavy-tail distribution of flights in human walks. A heavy-tailed distribution of flights is a signature feature of Levy walks. The data analysis based on GPS traces of human walks reveals that the sizes of a few extremely large hot spots are dominating the mean size of hot spots and they cause the bursty (i.e., long-range dependent) dispersion of visit points where people make a stop. Bursty visit points cause the characteristic distance among visit points to have a heavy-tail distribution. On top of bursty visit points, humans perform a distance-optimizing algorithm to plan their trips much like a heuristic to the traveling salesman problem. These factors in combination make human walks to have a heavy-tailed flight distribution. The above findings enable the construction of a simple human mobility model that taking as input the degree of burstiness in visit point dispersion, can naturally emulate hot spots as well as a heavy-tail flight distribution, both known to be important in measuring the realistic performance of mobile networks.

1. INTRODUCTION

Rhee et al. [25][24] show that human walk patterns contain similar statistical features found in Levy walks which biologists have observed from the mobility patterns of animals like spider monkeys [23] and seabirds [28]. Levy walks (LW) are characterized by a power distribution of flights where a flight is the distance that a walker travels without making a pause or a directional change. [25] shows that human walks are statistically and fundamentally different from walks generated from commonly used mobility models such as random way point (RWP), random direction and Brownian motion (BM) whose flight distributions have a short tail. It shows that the flight and pause time distributions of human walks have a strong heavy tail tendency, and further, while the mobility of humans is super-diffusive, its diffusivity falls somewhere between the diffusivity of RWP and that of BM. These features of human walks lead to performance characteristics of network protocols that form a middle ground between those seen from BM and RWP. For instance, the heavy-tail flight distribution of LW induces on average much longer inter-contact times (ICTs) than RWP, but shorter than BM. That is, Levy walkers meet much more often than Brownian walkers, but much less often than RWP walkers. This results in much longer DTN (delay-tolerant network) routing performance than RWP, but much shorter DTN routing performance than BW. Thus, many existing DTN routing performance studies using RWP are significantly over-estimating the DTN routing performance.

Unfortunately, human walks are not LW, though they have some similar statistical features, since unlike LW people hardly move randomly. They move with and within numerous contexts such as buildings, classrooms, market places, shopping malls, restaurants, tourist attractions, schedules and appointments. Random walk models are too simplistic to represent these man-made contexts. Obviously not all contexts are relevant for mobility modeling. Then, what contexts make human walks have such a heavy tail tendency? Finding fundamental causes of these LW features in human walks helps identify the contexts inducing such features. From these "seed" contexts, human walks can be naturally generated without altering its distinctive statistical features. This way of generating mobility traces has the following unique advantages over existing mobility models.

1. **Generating natural hot spots.** Since it represents important contexts where people walk in and out and also gather around, it naturally forms swarms or hot spots where many people "meet". Representing hot spots naturally captures the realistic inter-contact properties of humans. In contrast, random walks cannot represent hot spots and thus are limited in capturing such properties.

2. **Generating natural Heavy-tail flights.** It naturally represents the similar statistical patterns as LW, but without forcing walkers to "jump" to random locations in order to make heavy-tail flights. As discussed above, heavy-tail flights induce a power-law ICT distribution, an important statistical characteristic for DTN routing. Artificially placing hot

spots as done in many other mobility models [5] [17] incorporating hot spots does not necessarily create such features.

3. **Simplifying modeling.** One way to faithfully represent man-made contexts is to model detailed information about many man-made contexts such as schedules, appointments, hot spot locations and sizes, human activities around hot spots, as done in detailed simulation (e.g., [16]). But such modeling is too costly in terms of time and resource. In contrast, modeling seed contexts that cause heavy-tail flights limits the amount of efforts required for constructing such a mobility model without sacrificing important features of human mobility.

This paper provides reasonable answers to the causes of heavy-tail flights in human walk patterns. Using the GPS traces used in [25] which are taken from about 100 participants in two university campuses, Disney World, New York City and a state fair involving about 200 daily traces, we make the following observations.

First, daily walk traces contain highly bursty *visit points* which are defined as the places where the participant makes a stop. They form swarms of highly varying sizes and their swarm sizes show that their normalized variance over increasing ranges of aggregation decays very slowly, with Hurst values ranging from 0.8 to 0.6, and their distribution shows a heavy-tail tendency. A surprising finding is that the aggregation of visit points of all the participants within the same site is also very bursty and long-range dependent. This is surprising because participants are chosen arbitrarily and have very little in common except that their primary walk-about areas are confined to the same area. This implies that people tend to visit locations that other people also visit, and popular places tend to be extremely popular whose popularity dominates the mean. This finding is important because it reveals the statistical features of hot-spots where people swarm, and it is known that hot-spots strongly influence wireless routing performance [2].

Second, from the daily traces, we find that participants plan their trip to their visit points to optimize the total sum of flights using a higher order heuristic function of distance. This way of planning is similar to heuristics to a combinatorial NP-hard problem called *traveling salesman* (TS) [4] which minimizes the total travel distance while visiting all the input locations from a starting location, and finally coming back to the original location. This observation coincides with the least-action theory of Maupertuis [6]. We find that the heuristics using a function $1/d^a$ where d is the distance from a current point to its next point and a is a positive constant, gives extremely good matching to the original GPS traces with less than 1 to 11% error margin. If a is infinite, then it uses the heuristic algorithm called nearest neighbor first (NNF)[26] [22] for TS. In most cases, we find a between 1 and 3 provides very good matching.

From these observations, we identify the seed contexts for mobility modeling to be bursty visit points. This leads to a construction of simple contexts from which both hot-spots and heavy-tail flight distributions of human walks can be naturally produced from the seed context. We generate bursty hot spots by emulating the burstiness of visit points measured from real traces. Then for a daily trip, each mobile node chooses its visit points from these points also in a bursty manner. Synthetic walk traces are generated by visiting these points using a function of $1/d^a$ where we vary a from 1 to 3. We call our model *bursty spot model* (BSM).

We apply BSM to the evaluation of several DTN routing protocols and verify that our model generates heavy-tail flight. The emulation of hot spots reduces DTN routing performance substantially from popular random walk models– an order-of-magnitude less than LW and even 2 to 3 times less than RWP. This result indicates that the statistical features of hot spots and flights make significant impacts on the performance of routing in mobile networks and realistic representations of them are important.

The remainder of this paper is organized as follows. Section 2 discusses related work, Section 3 describes the human walk traces from [25] used for our data analysis, Section 4 presents the results of our data analysis, Section 5 describes BSM, Section 6 presents the DTN routing performance results using BSM and Section 7 concludes this paper.

2. RELATED WORK

A mobility model is placed somewhere between a pure random model and a target real environment. Some are close to a pure random model like as BM, RWP, random direction mobility model, and so on, others try to be realistic. Human mobility is affected by geographical constraints and his intention.

Jardosh et al. [13] incorporate obstacles in their mobility model called obstacle mobility (OM) to emulate more realistic navigations of humans around obstacles. They approximate obstacles as polygons, and place them in a simulation terrain. Pathways are created using Voronoi diagrams. The routing performance using OM is significantly different from that using RWP. The location-based preferences and hot spots have also been modeled using a weighted way point model [14] and a Markovian waypoint mobility model [12]. These models do not consider realistic statistical features of human mobility.

Hong et al. [11] model groups of users that move together, but the model does not consider realistic representations of groups. Based on Albert and Barabassi's

preferential attachment theory, Herrmann [10] first incorporate a power-law social inter-action model into human mobility. Musolesi et al. [21] extend the work of Herrmann by incorporating geographical movements of groups. But these models do not consider the geographical positions of meetings and hot spots.

Borrel et al. [5] present a mobility model using preferential attachment. In the model, attractors and people dynamically arrive to the simulation terrain according to a Poisson process and stay for a random duration. An attractor can be considered as a landmark or a hot spot. Each individual chooses a destination attractor with a probability proportional to its attractiveness which is proportional to the number of attracted individuals and inversely proportional to distance between the individual and the attractor. Since the model does not consider bursty placements of hot spots, the generated flights do not follow a heavy-tail distribution. Lim et al. [17] also use preferential attachment. It first divides the simulation area into several subareas and use them as hot spots. Initially, users are assigned to a subarea using preferential attachment, then they choose a next subarea according to attractiveness proportional to $(k+1)^{\alpha}$, where k and α are the number of users in the subarea and the clustering exponent, respectively. The strength of scale-free phenomena can be adjusted according to α. This model also has the same problem as Borrel's in that it does not create a heavy-tail distribution of flights because attractors are distributed uniformly and no distance consideration is given for trip planning.

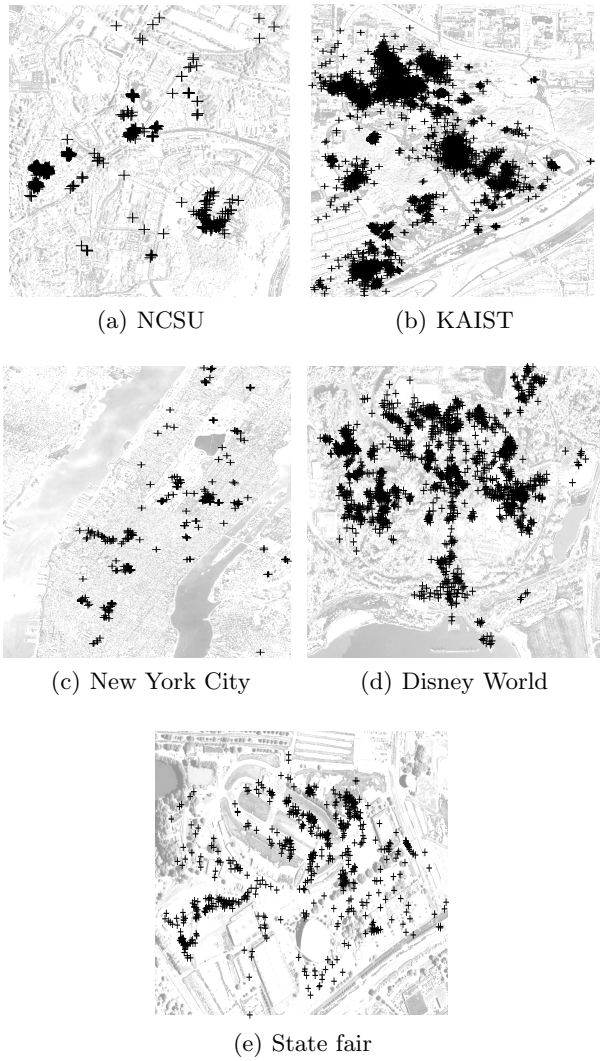
There have been a few studies to analyze mobility patterns of users using wireless devices [9] [20]. These studies use periodic log data or event log data associated with mobile devices at IEEE 802.11 access points (APs). Kim et al. [16] estimate the locations and movement paths of users from the sets of AP log data. Based on the estimated information on locations, pause time, and velocity, hot spot regions and the transition probability for moving from one hot spot to another are extracted. This information is used to construct a mobility model. This model can generate a fairly realistic and detailed representation of human mobility. But it requires a considerable amount of effort to generate the mobility model because hot spot locations and transition probability between hot spots must be given as input (instead of being generated).

3. HUMAN WALK TRACES

Five sites are chosen for collecting human mobility traces. These are two university campuses (NCSU and KAIST), New York City, Disney World, and North Carolina State Fair. The total number of traces from these sites is over 150 daily traces. Garmin GPS 60CSx handheld receivers are used for data collection which are

Site (# of participants)	# of traces	Duration (avg) [hour]	# of visit pts (avg)	X length of site [meter]	Y length of site [meter]
KAIST (34)	76	1445.15 (19.02)	10598 (139.45)	10256.28	18650.72
NCSU (20)	31	310.62 (10.02)	3316 (106.97)	2586.85	2347.01
State fair (18)	18	47.01 (2.61)	691 (38.39)	1141.02	995.75
Disney World (18)	38	378.08 (9.95)	4085 (107.50)	8214.56	9446.70
New York (10)	32	293.21 (9.16)	1105 (34.53)	31432.72	18900.42

Table 1: Statistics of collected mobility traces from five sites.



(a) NCSU (b) KAIST

(c) New York City (d) Disney World

(e) State fair

Figure 1: Visit points registered in our traces. Visit points are marked by '+'.

WAAS (Wide Area Augmentation System) capable with a position accuracy of better than three meters 95 percent of the time, in North America [1]. Occasionally, track information has discontinuity mainly when bear-

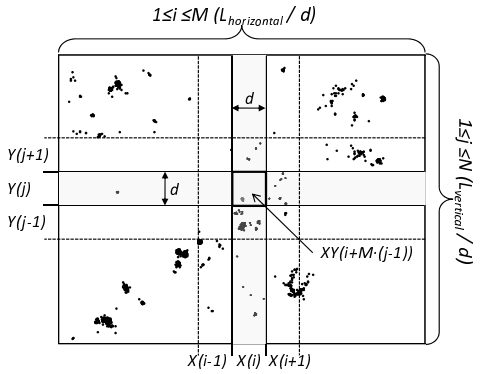


Figure 2: Measuring aggregated variance of visit points aggregated from all walk traces. We divide the area by non-overlapping d by d squares, and count the number of visit points registered in each square and then normalize the sampled count by the size of the unit square. We compute the normalized variance as we increase d.

ers move indoor where GPS signals cannot be received. The GPS receivers take a reading of their current positions at every 10 seconds and record them into a daily track log. The summary of daily traces is shown in Table 1. The radius of each trace is a half of the maximum distance that a participant travels during a day.

20 participants in NCSU were randomly selected from the students who took a course in the computer science department. Every week, 2 or 3 randomly chosen students carried the GPS receivers. The KAIST traces are taken by 34 students who live in a campus dormitory. Since the participants in NCSU and KAIST occasionally moved outside their campuses, we use only those logs recorded within a radius of 10 km from the center of each campus. The New York City traces were obtained from 10 volunteers living in Manhattan or its vicinity. Most of the participants have offices in Manhattan. Their track logs contain relatively long distance travels because of their long commuting paths. Their means of travel include subway trains, buses and mostly walking. The State fair track logs were collected from 18 volunteers who visited a local state fair that includes many street arcades, small street food stands and showcases. The site is completely outdoor and is smallest among all the sites. Each participant in the State fair scenario spent less than three hours in the site. The Disney World traces were obtained from 18 volunteers who spent their Thanksgiving or Christmas holidays in Disney World, Florida, USA. For our study, we use only the track logs from the inside of the theme parks. The participants mainly walked in the parks and occasionally rode trolleys.

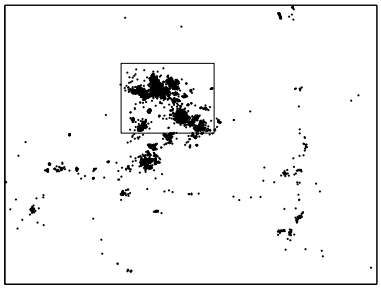
4. MEASUREMENT STUDY

In this section, we examine the causes of heavy-tail distributions of flights in human walks. We conjecture that they are highly related to the locations of destinations where people choose to walk to. To confirm this, we define a *visit point* to be the GPS location where a participant stays more than 30 seconds within a circle of 5 meter radius of that location. For each site, we plot the visit points registered by every walk trace of that site. We call these points *aggregated visit points*. Figure 1 shows the aggregated visit point of each site.

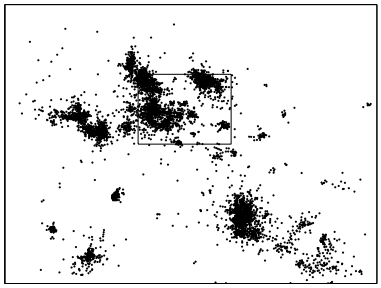
To measure the burstiness in the dispersion of visit points, we divide the site map into a grid of unit squares (initially of 5 by 5 meters). We count all the visit points within each square and then normalize the count by the area of the square. We measure the variance in these normalized count samples and call it *aggregated variance*. Figure 2 illustrates the method. If there exists a long-range dependency in the samples, the aggregated variance should not decay faster than -1 in a log-log scale as we increase the size of the square. To see this, we plot aggregated variance in a log-log scale as we increase the square size and measure its absolute slope β. The *Hurst* parameter of the samples is $1 - \beta/2$. The sample data are said to be *bursty* or *long-range dependent* (and therefore, self-similar) if the Hurst parameter is in between 0.5 and 1. Aggregated variance can also be computed over one dimension by mapping visit points to X or Y axis of the map. In this case, we use a line instead of a square.

4.1 Bursty aggregated visit points

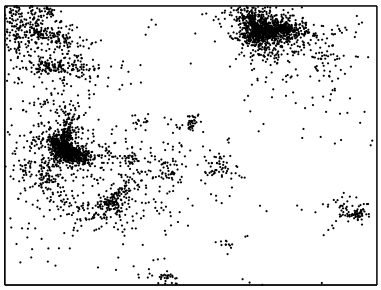
The bursty dispersion of visit points (or simply bursty visit points) implies that people tend to swarm near to a few popular locations and their popularity measured by the number of visit points within the swarms of visit points formed around the locations shows high burstiness: the popular locations tend to be very popular while most other areas are not. Figure 3 plots the aggregated visit points of KAIST while we zoom in to smaller areas (denoted by small boxes) in the map. The patterns of swarming distinctively appear similar independent of their zoom resolution (or scale). Figure 4 shows the Hurst parameter measured from the aggregated visit points of KAIST. These values show a very strong long-range dependency with a Hurst value larger than 0.8. Figure 5 plots the Hurst values measured from all the sites with their 95% confidence intervals. All the sites except NYC show a high degree of burstiness while the NYC traces show only slight burstiness. This outlier is, we conjecture, due to the very small number of participants relative to the size of the area and the number of registered visit points are relatively small. Except NYC, the burstiness of the visit points is evident independent of their site locations although the degree of burstiness may vary from one site to another.



(a) 4800m×1200m



(b) 1200m×600m

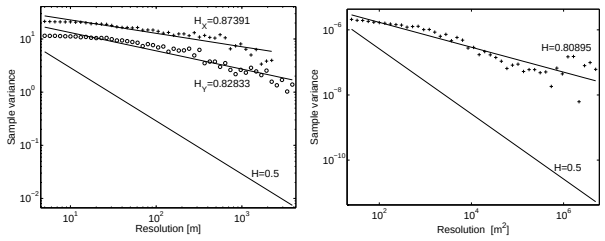


(c) 300m×300m

Figure 3: It shows the self-similar nature of the dispersion of visit points in the KAIST trace. At different scales, the dispersion of visit points looks similarly bursty.

4.2 Bursty hot spots

Figure 1 shows that aggregated visit points are clustered to form hot spots. This is because different people may visit similar locations. A hot spot is defined to be a cluster of visit points that are connected to each other by the transitive closure of the connected relation. We say that two visit points are *connected* if they are within a predefined radio range. The size of a hot spot is the number of visit points within that hot spot. Figure 6 shows a clear pattern of heavy-tail distributions in hot spot sizes over several orders of magnitudes. It is a natural consequence of the burstiness in aggregated visit points shown in Section 4.1. This phenomenon coincides with the preferential attachment theory of Albert and Barabassi [3] where popular places become more popular to form a heavy tail distribution of popularity.



(a) 1-D stripe (b) 2-D grid

Figure 4: Hurst parameter estimation of visit points registered in all KAIST traces.

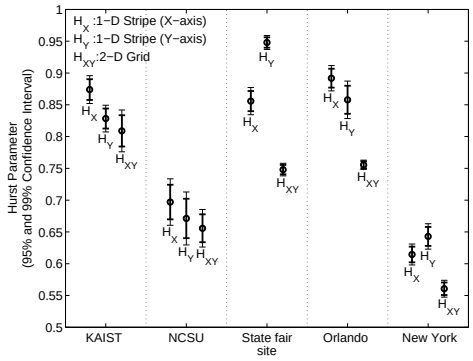


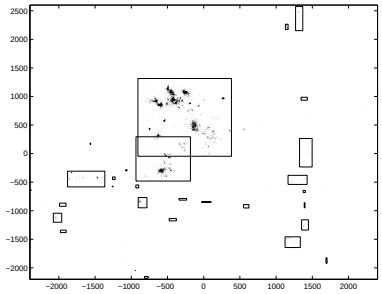
Figure 5: Hurst parameter values of visit points in each site map. All show Hurst values higher than 0.6 except NYC.

4.3 Bursty individual traces

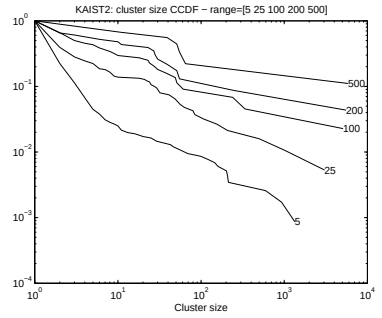
We also observe that the visit points registered in each individual trace are bursty. For each trace, we perform the aggregated variance test on its visit points. Figure 7 shows the Hurst parameter values of individual traces from the five sites. Their H values are slightly less than those from the aggregated visit points shown in Figure 5, confirming that burstiness gets intensified as individually bursty traces are superimposed together.

Does the burstiness of individual traces come from the burstiness of hot spots or vice versa? Our data analysis suggests the former. To see this, we randomly pick a subset of visit points from the aggregated visit points of each site without any bias to locations. We find that these random subsets also show burstiness similar to that in individual traces. Figure 8 shows the Hurst values of visit points randomly taken from the aggregated visit points of KAIST. All 76 traces show a high degree of burstiness. However, aggregating any random, but bursty visit points does not necessarily result in bursty hot spots. Our data shows that there appears clear gravity to spots that allow visit points to form clusters and that the burstiness in individual traces arises from that of hot spots, but not vice versa.

Table 2 shows the statistics from individual traces

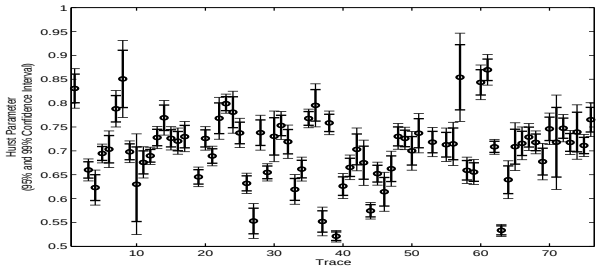


(a) Hot spots

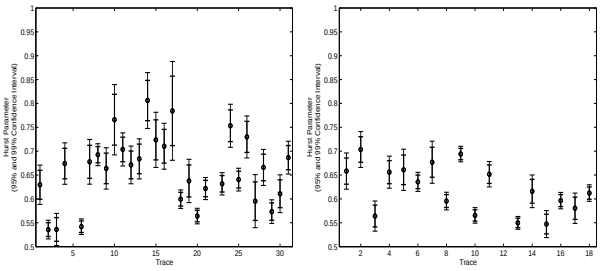


(b) CCDF

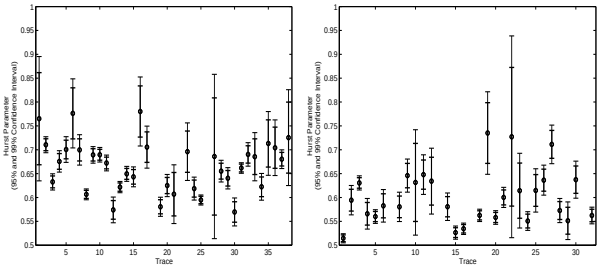
Figure 6: The hot spots in KAIST formed with 100 meter range and CCDFs of their sizes for different range values. The boxes in (a) represent the tightest rectangles enclosing all the visit points in the same clusters.



(a) KAIST



(b) NCSU (c) State fair



(d) Orlando (e) New York City

Figure 7: Hurst parameter values of visit points registered in each individual trace of the five sites. All traces show a tendency of self-similarity.

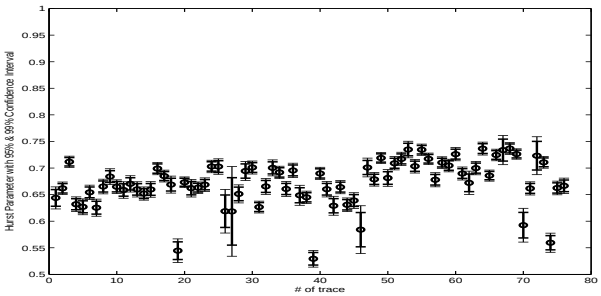


Figure 8: Hurst parameter values of visit points randomly taken from the aggregated visit points of KAIST. This indicates that the bursty visit points seen in individual traces arise from bursty hot spots (or heavy-tail distribution of hot spots).

	NCSU	KAIST	NYC	DW	SF
Site	43	98	117	52	6
Trace	4.55	3.66	6.13	3.34	1.67

Table 2: The number of hot spots (clusters) in each site with a radio range of 100 meters and the average number of hot spots visited in each trace.

on the number of hot spots (when a radio range is 100 meters) in each site and the average number of hot spots that each trace visits. On average, each trace visits about 5 to 10% of hot spots present in that site.

4.4 Characteristic distance

Flights are created on top of visit points. Depending on the choices of next visit points, the traces may have different flight distributions. What aspect of mobility causes a heavy-tail flight distribution? We find that this is in part due to heavy-tail distributions of distance between clusters of visit points at different scales. The intuition is as follows. Long flights are created when a person leaves one hot spot to get to another hot spot. Since the dispersion of visit points is long-range dependent, the probability that there exists a neighboring hot spot (swarm or cluster of visit points) at the increasing scales of range, decays very slowly even if clusters are coalesced together as the range increases. If flights are created by moving from one cluster to another, flights will also likely have a heavy tail distribution because a longer distance between clusters always exists.

Mendelbrot [19] made a related observation about fractal points (or the dispersion of points in a bursty manner)that when fractal points are dispersed over 1-D, their gaps (the distance between two neighboring points) have a power-law distribution. This observation has been extended to multi-dimensional gaps [7] where fractal points are dispersed in multi-dimensional spaces, their multi-dimensional gaps (or voids) also have a power-law distribution. To show this, Delaunay triangulation is used to measure gaps. Delaunay triangles are defined to be triangles formed over a set of points in a plane in which no points are inside of the circumcircle of any triangles.

Figures 9 show Delaunay triangles drawn on top of the aggregated visit points of each site and the CCDF of the lengths of the lines forming the triangles. These CCDFs are showing the same heavy tail distributions found from real human walk traces. These CCDFs and the CCDFs of flights from real traces are surprisingly similar. This strongly implies that humans tend to choose as their flights the lines used in Delaunay triangulation. Since Delaunay triangles are formed between only "natural" neighbors in a planar graph, This match of CCDFs indicates that humans tend to visit all neighboring points first before they make a long-jump to neighboring clusters of visit points. This phenomenon is further verified in the next section.

4.5 Least Action Trip Planning

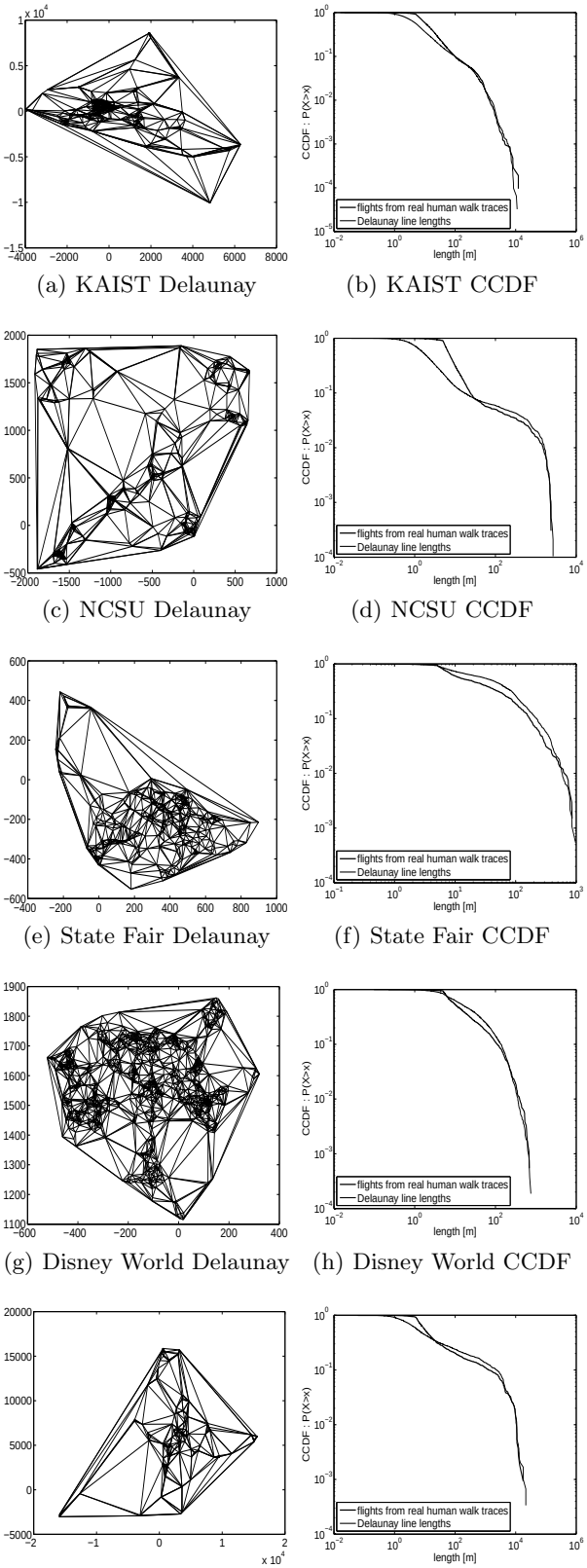
Given a set of visit points, how do humans plan their trips around these visit points to form a heavy-tail distribution of flights? Flights are highly influenced by the ways that humans choose the next visit points from their current visit points. In this decision, many factors play a role. Since every person may have different factors, cost functions, personal situations and personal tendency, it is almost impossible to derive one algorithm that can apply to all cases. The results from Delaunay triangulation strongly indicate that people visit nearby visit points first before jumping to visit points in a different cluster (or hot spot). In a way, this pattern minimizes the total amount of distance that a person travels, implying that distance is a stronger determinant in this decision.

Maupertuis' principle of least action [6] provides some clues to this phenomenon: humans tend to make actions that require the least amount of effort. Helbin et al. [**?**] also show that when people make trails in a park, they use a cost function minimizing discomfort in moving from one place to another. This discomfort can be translated into the distance as well as the condition of roads (i.e., whether it is paved or not). To study the influence of distance in choosing paths by humans, we run simple simulation mimicking a distance-based path choosing mechanism of humans when a set of visit points V is given. The simulation runs using the following path selection algorithm.

Path selection algorithm (least action trip planning). At the current position i in V, the probability that a next position j is chosen is computed as $\frac{1/d_{ij}^a}{\sum_{\text{for all } k \in V-V'} 1/d_{ik}^a}$ where d_{ij} is the Euclidean distance from i to j, a is a fixed floating number constant within 0 to infinity, and V' is the set of positions in V that have been visited so far. Based on this probability, a next position is randomly chosen from V'.

From the algorithm, if a is set to zero, the next visited point is completely independent of the distance to that next point, much like RWP, and if a is infinity, then an unvisited point with the shortest distance to p is chosen as the next position and this algorithm is in fact the nearest neighbor first (NNF) heuristic used for the traveling sales problem [26] [22]. As a increases, people place more weights on distance in their path selection decisions.

We perform the simulation on top of visit points taken in each individual traces using the algorithm and the results are given in Figure 10. We also compare the results with those from Levy Walks (LW) and RWP. With a set to 1.5 or 3, the difference in the sums of flights from



(a) KAIST Delaunay (b) KAIST CCDF

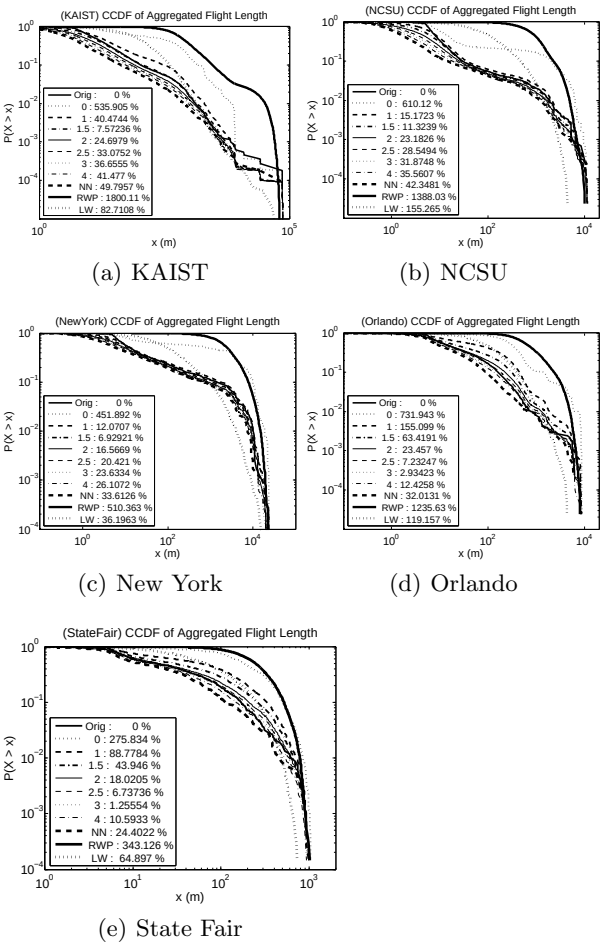
(c) NCSU Delaunay (d) NCSU CCDF

(e) State Fair Delaunay (f) State Fair CCDF

(g) Disney World Delaunay (h) Disney World CCDF

(i) New York City Delaunay (j) New York City CCDF

Figure 9: Delaunay triangulation of visit points, and the CCDF of the lengths of lines formed by the triangulation on top of the visit points extracted from individual traces.



(a) KAIST (b) NCSU

(c) New York (d) Orlando

(e) State Fair

Figure 10: The CCDF of aggregated flights formed by the path selection algorithm with various values of a. The legend inside each figure denotes the value of a and its corresponding errors in flight sums.

the simulation and the real traces is always between 1% to 11%, and the CCDFs of flights also match extremely well, indicating they have very similar means and variances. Compared to LW and RWP, our model produces far better matches.

We can interpret the above result as follows. $a = 3$ produces a good match to flight distributions taken from theme-parks (Disney world and State Fair) as humans plan their travel by the distance to each attraction in theme-parks and try to minimize the total traveling distance as most try to visit as many attractions as possible within a given time, and with a 1.5, other factors play slightly bigger roles (e.g., with time critical events like predetermined meetings and class schedules, people have to travel to the meeting location irrespective of its distance). Thus this case fits better to the campus scenarios. Even in these scenarios, distance has a strong influence.

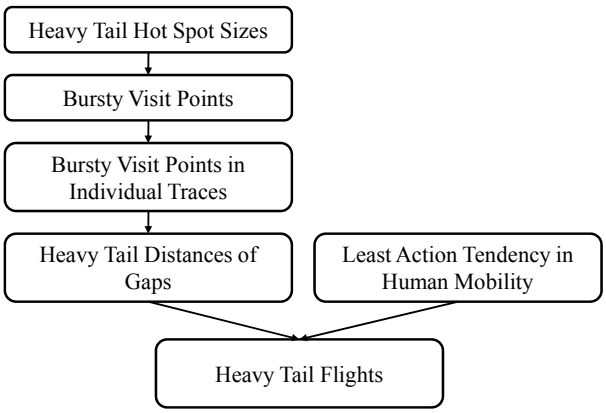


Figure 11: The summary of our finding from the data analysis of human walk traces.

This result permits the logical link between the bursty visit points in human traces and the heavy-tail flight distributions. In Section 4.4, we show that the bursty visit points induce the heavy tail distributions of the lengths of the lines formed by Delaulay triangles of visit points. As people give more importance to distance when making path selections, the nearest unvisited neighboring point is more likely chosen and that path is likely part of a Delaunay triangle. Since the lengths of the lines forming the Delaunay triangles have a heavy-tail distribution, it is quite natural that flights also have the same tendency.

4.6 Summary of measurement study

In summary, our measurement study provides a plausible explanation for the causes of Levy walk patterns in human walks. (1) Section 4.2 shows that human visit points form clusters whose sizes are bursty (i.e., long-range dependent) and also has a heavy-tail distribution, (2) Section 4.3 shows that visit points found in individual traces are also dispersed in a bursty manner, and this tendency arises from the bursty hot spots, (3) Section 4.4 shows that visit points dispersed in a bursty manner in an area induces a heavy-tail distribution of the distance of "gaps" between visit points and (4) Section 4.5 shows that distance-based path selection on top of bursty visit points results in a heavy-tail distribution of flights observed in real walk traces. Figure 11 summarizes our data analysis results.

5. BURSTY SPOT MODEL

Our data analysis indicates that bursty visit points and distance-based trip planning over those visit points are the keys in generating a heavy-tail flight distribution observed in human walk traces. This observation greatly simplifies the construction of a human walk model that can emulate both hot spot and flight statistics found in real traces. This section discusses the construction of such a model called Bursty Spot Model (BSM).

We generate hot spots by dispersing visit points in a bursty manner in a given area S. This dispersion creates a synthetic map G containing the locations of a fixed number of visit points given as input. The degree of the burstiness must match closely that from a real map T produced from human walk traces. To achieve this, we first divide an input area A into N square segments of an equal size and then we distribute n visit points given also as input over to the N segments while ensuring that the normalized variance in the numbers of points assigned to each segment is matched to the input variance R. The *normalized variance* is defined to be $Var(\frac{X}{E[X]})$ where X is the number of visit points assigned to a segment. It can be proven that normalized variance is equivalent to aggregated variance in Secton 4.1. Note that this is different from the aggregated variance in Section 4 where the sample is divided by the area. We recursively apply the same technique: for each segment i, A is set to i, the number of points assigned to i as n, and $\bar{v}_l^T/\bar{v}_{l-1}^G - 1$ as R. $\bar{v}_l^T$ and $\bar{v}_{l-1}^G$ are respectively defined as follows. l is the resolution level (scale) and initially set to one. For each recursive call, l is incremented by one. The structure of segments divided into l levels looks exactly like a full N-ary tree with height l. The number of segments at level l is N^l. $\bar{v}_l^T$ is the normalized variance at level l measured from T where a sample X is the number of visit points in a segment at level l in T. $\bar{v}_{l-1}^G$ is the normalized variance measured from the synthetic trace G generated at the previous level (i.e., level $l-1$). Initially, A is set to S, n is the total number of visit points in T, and $\bar{v}_0^G$ is set to 1. The maximum number of levels is fixed to some constant (typically we use 9).

The rationale for this scheme starts from the notion that by emulating the same variance observed at each scale from the original trace, we can emulate the self-similarity of the original trace. However, matching the variances in the synthetic trace and the original trace is not straightforward. At level 1, matching the variance is simple since the same number of visit points as in T must be distributed. But at a higher level, the number of visit points assigned to each segment may now be different. Since we need to match only the variance but not the mean, we normalize each sample (X) by the mean. Our goal is to ensure $\bar{v}_l^T = \bar{v}_l^G$. The problem gets a little complicated since we need to achieve this goal by adjusting the variance at each segment independently. This problem can be mapped into a line segmenting problem where a line with length 1 is divided into N line segments whose length distribution has a normalized variance $\bar{v}_1^T$ and each segment i is then again divided into N segments whose distribution R of lengths divided by the length of segment i is set to produce the variance

of lengths of line segments measured from level 2, i.e., $\bar{v}_2^T$. This process continues until we reach the final level.

Now the problem is reduced to a problem of finding such an R for each level. To recap the problem, at level $l-1$, there are N^{l-1} line segments with a distribution $A = [p_1, p_2, \ldots, p_{N^{l-1}}]$ and the normalized variance is $\bar{v}_{l-1}^G$ and each segment p_i is divided into another N segments with the following ratio distribution $B = [q_1, q_2, q_3, \ldots, q_N]$ whose variance is R and sum is also one. Thus, the resulting line segments have a distribution C which contains the line segments of lengths $p_i q_1, p_i q_2, \ldots, p_i q_N$ for all i's. Let X, Y and Z be random variables that have the same distributions as A, B and C respectively. Note that $E[Z]$ is $1/N^l$ since the total length is one and there are N^l segments.

$$\begin{aligned} \bar{v}_l^G &= \frac{E[Z^2]}{E[Z]^2} - 1 \\ &= (4^{2l})E[Z^2] - 1 \\ &= (4^l)\sum Z^2 - 1 \\ &= (4^2 E[Y^2])(4^{2(l-1)}E[X^2]) - 1 \\ &= (R+1)(\bar{v}_{l-1}^G + 1) - 1 \end{aligned} \tag{1}$$

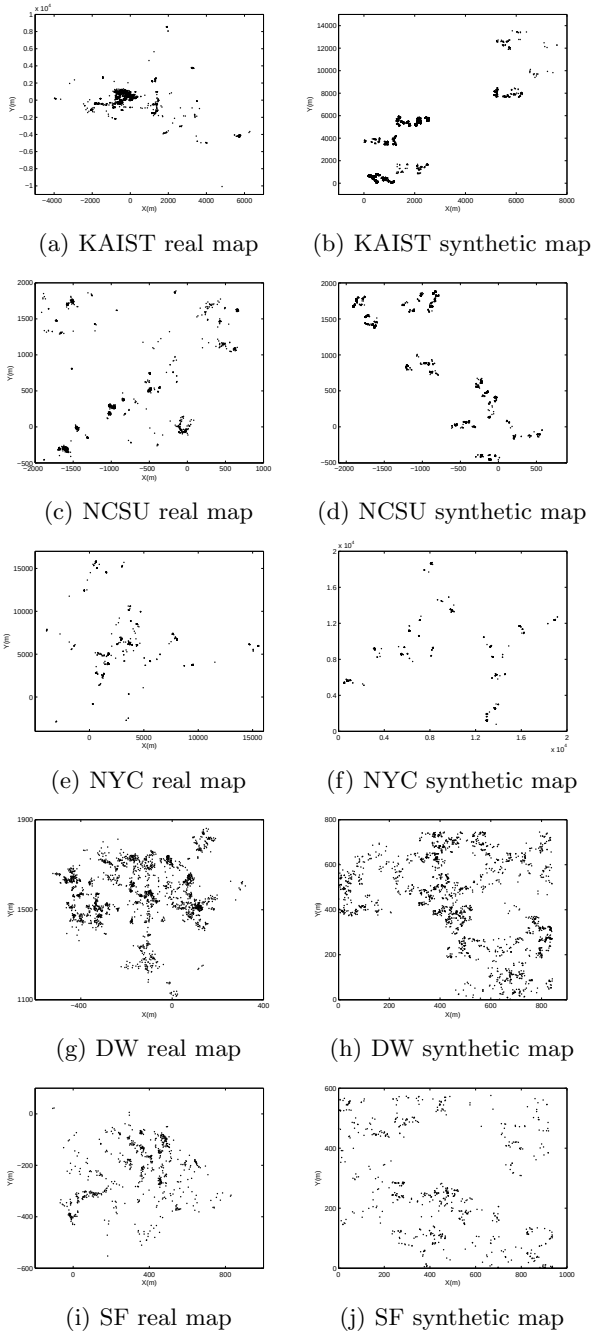
Therefore, if $R = \bar{v}_l^T / \bar{v}_{l-1}^G - 1$, then we have $\bar{v}_l^T = \bar{v}_l^G$. This method takes as input the total number of visit points, $\bar{v}_l^T$ for each level l, and the area S.

Note that normalized variance can be plotted as a straight line in a log-log variance vs. scale plot as shown in Figureẽffig:KAIST-hurst. Thus, instead of providing a set of normalized variance as input, we can simply provide a starting variance (the variance at the first level) and a power-law slope (between 0.5 and 1) as input to the hot spot generator.

Figure 12 shows the sample synthetic trace generated from the above method from the statistics measured from each site.

After we generate a map of visit points, each node selects a subset of visit points in the map and plans a trip over the subset using the path selection algorithm described in Section 4.5. From Table 2, we find that each individual walk trace visits about 5% to 10% of hot spots in the map. Thus, after we form hot spots with a fixed range using the technique described in Section 4.2, the subset V of visit points to visit in a single synthetic walk trace being generated is selected as follows. (1) We randomly (i.e., with uniform distribution) select 5 to 10 % of hot spots to visit while assigning proportionally more weights for the selection of a bigger hot spot, and then (2) from the visit points belonging to the selected hot spots, again randomly select V. The number of visit points to select for V is given from user input.

Figure 13 describes the scheme for generating BSM



(a) KAIST real map (b) KAIST synthetic map

(c) NCSU real map (d) NCSU synthetic map

(e) NYC real map (f) NYC synthetic map

(g) DW real map (h) DW synthetic map

(i) SF real map (j) SF synthetic map

Figure 12: Maps containing visit points registered in real traces are compared to synthetically generated maps.

traces. All the components used in this scheme except the generations of hot spots and per-trace visit points are discussed in Sections 4.3 and 4.5.

Figure 14 measures the CCDF of the flights in the synthetic walk traces of BSM constructed using the input values extracted from real traces of each site (i.e., the same number of visit points and the same normal-

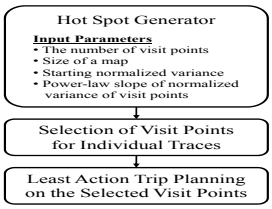
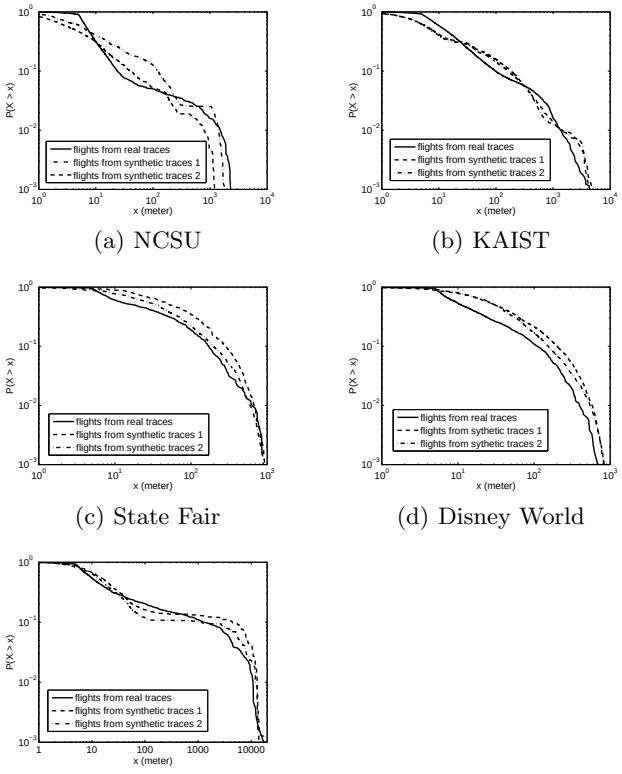


Figure 13: The schematic diagram of Bursty spot model.



(a) NCSU (b) KAIST

(c) State Fair (d) Disney World

(e) New York City

Figure 14: The CCDFs of flights from synthetic maps of visit points generated by BSM using normalized variance values taken from real traces. They match very well with those from their corresponding real traces.

ized variance values). It is compared to those from real walk traces in the same sites. They all show a very close match, verifying that BSM reliably generates realistic flight distributions from the given input.

6. ROUTING PERFORMANCE

In this section, we will examine the impact of the features of our mobility model on the performance of human driven DTNs. We have the following setup for the simulation.

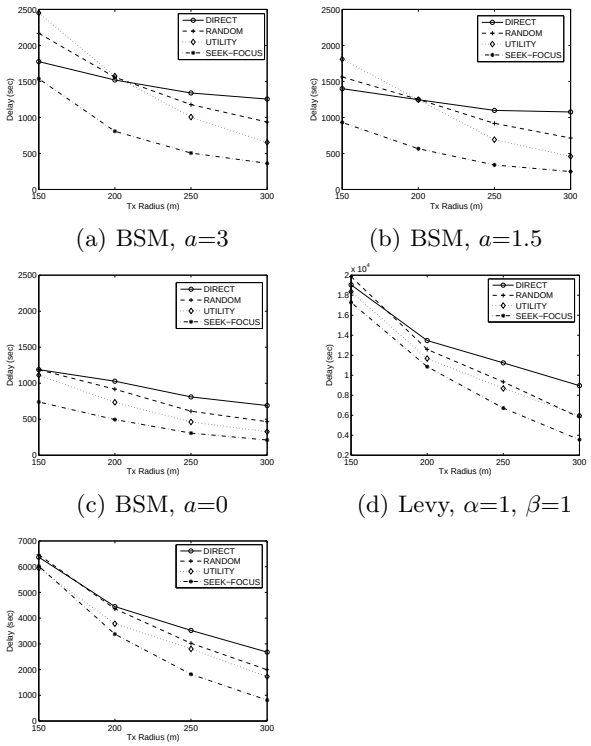
6.1 Simulation Setup

We simulate 250 hours of human walks of 50 persons. We generate five different simulation sites similar to the five sites where the human walk traces are taken. Based on information in Table 1, we fix the size of the simulation areas and the number of visit points (applied only to BSM). We also fix the slope of normalized variances to the average β values of aggregated variance measured from each site. Their values can be deduced from Figure 5. For BSM, we vary a from 0 to 3. We denote BSM(k) to indicate BSM simulation with $a = k$. Every mobility model has the same truncated power-law distribution of pause times (its slope is one and its maximum pause time is 28 hours) – the distribution is obtained from real walk traces [25]. For LW, we use the truncated power-law distribution of flights with slope 1 and the maximum flight length 2.5km. During the simulation, the contact information is checked at every 1 minute. The initial position of every person is selected randomly from visit points and all mobility models have the same starting points. We discard the first 50 hours of simulation results to avoid any transient effects. The speed of every user is set to 1 m/s for simplicity. Unless specified otherwise, we set the transmission range of mobile devices to 250m which is the typical value of WiFi.

We run four of the most widely studied DTN routing protocols over the generated mobility traces. The simplest one is *Direct Transmission* where a message is transferred only when a source node finds its destination node in its radio range [8]. It is a very trivial algorithm but is used as a baseline. *Randomized Routing* [27] allows the holder of a message to send the message to another node in a radio range with probability p which is between 0 and 1. In *Utility-based Routing Without Transitivity* [15] [18] [27], each node forwards its message to the neighboring node with the maximum utility value among all of its neighbors. The utility function is set to be the age of the last encounter with a particular node. Thus, a node forwards to a neighbor that has met the destination most recently among all its neighbors. *Seek and Focus* [27] is a hybrid approach rhat selectively uses randomized routing and utility-based routing. If the utility value (the time duration after the last encounter with the destination) is less than a threshold T_h, then utility-based routing is used. Otherwise, randomized routing is used. We fix T_h to 500 seconds.

6.2 Performance results

We run the four DTN routing algorithms in the NCSU site. Figure 15 shows the routing delay. In this simulation, for each run, we generate 300 messages with random source and destination pairs and we wait until 200 messages are delivered to their destinations before

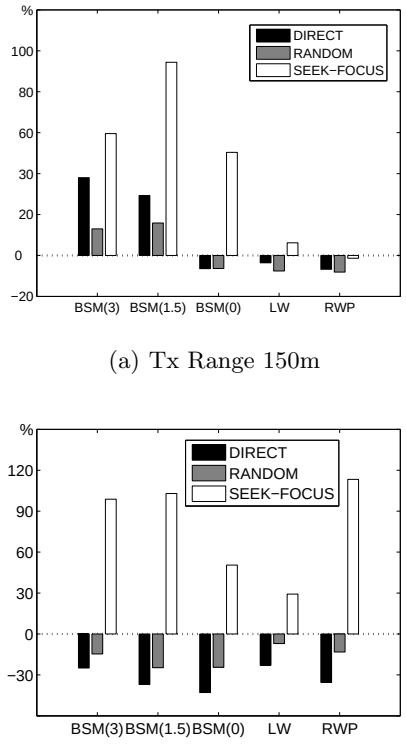


(a) BSM, a=3 (b) BSM, a=1.5

(c) BSM, a=0 (d) Levy, α=1, β=1

(e) RWP

Figure 15: Average routing delays of four DTN routing protocols under various mobility models.



(a) Tx Range 150m

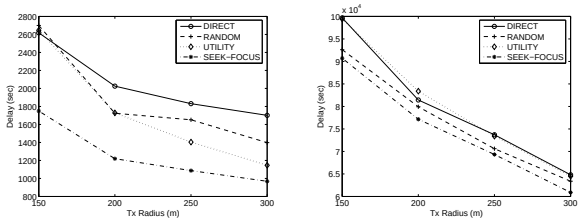
(b) Tx Range 250 m

Figure 16: The delay improvement ratio of various protocols over utility-based protocols under various transmission ranges.

measuring the routing delays. We repeat the test for 10 times for each setup. The figure shows the average routing delays for the delivery of 200 messages.

Several significant observations can be made about the difference in the routing performance when different routing protocols and mobility models are used. (1) Overall, the average routing performance of various protocols under BSM is much shorter than RWP and LW and among BSM, larger a values give longer routing delays. Our measurement study indicates that human walks have a within a range of [1:3]. So BSM(0) tends to overestimate the routing delays compared to BSM(3). (2) The staleness of utility information increases with a smaller transmission range. In BSM(3), utility-based routing performs much worse than randomized routing under 150m Tx range. However this phenomenon is not shown with random walk models: BSM(0), RWP and LW. This is because the impact of wrong or stale utility information on routing performance is well compensated by random walk mobility patterns of these models. Since human mobility is not completely random, it is natural that a "mistake" in routing using stale information should be penalized more in real scenarios. This is consistent with the results from BSM(1.5) and BSM(3). (3) The performance improvement of a hybrid routing protocol like Seek and Focus over utility-based and randomized protocols is much larger with BSM(3) than the other models. To illustrate this, we measure the delay improvement of different protocols over utility-based protocol in Figure 16. It shows that under 150m Tx range, in BSM(1.5) and BSM(3), Seek and Focus shows about 50 to 100% improvement while in the others, their improvement is much smaller. The performance gain gets reduced with a larger Tx range because of increased connectivity. Overall, the emulation of hot spots and heavy-tail flights is important in characterizing routing performance.

To see the effect of the degree of burstiness in the dispersion of visit points, we measure routing performance also on top of the synthetic maps constructed using the trace information of New York City and KAIST. The NYC traces show a very steep slope of the aggregated variance of visit points due to the relatively smaller number of participants and visit points compared to its size of the site. On the other hand, the KAIST traces exhibit much more burstiness than the NYC traces for the comparable size of the site. In this simulation, we use $a = 1.5$, the best matching a val-



(a) KAIST (b) NYC

Figure 17: The routing performance under the synthetic maps created from the KAIST and NYC data. NYC has a steeper aggregated variance of visit points than KAIST for approximately the same size of the area. The figure shows that the routing performance under the NYC map is much worse than under the KAIST map.

ues taken from Figure 10. Figure 17 show the result. The performance under the KAIST traces is an-order-of-magnitude better than under the NYC. In fact, the NYC traces have a much fewer number of visit points than KAIST, but in the simulation, we have the same number of walkers. Thus, they are more likely to visit the same visit points. But their performance is much worse. In contrast, KAIST has several big hot spots (due to its burstiness) where almost every walkers visit every day. This has made a big impact on the performance. Thus, the burstiness of hot spots is critical in realistic estimation of routing performance.

In summary, the performance of DTN routing is very sensitive to the burstiness in the distribution of hot spots and heavy-tail flight distributions. The BSM models are shown to effectively capture this sensitivity.

7. CONCLUSION

Humans never walk randomly. Nonetheless, many mobility models use random walks. Random walks completely lack in man-made contexts, and thus significantly distort any performance evaluation results performed using these models. Especially, hot spots where people meet and visit commonly are very important for correct estimation of mobile network performance. Existing work lacks in realistically representing hot spots and their statistical properties. Our paper addresses this issue by emulating the statistical patterns of hot spots, namely highly bursty natures of hot spot dispersion and sizes. We found this is one of the causes for a heavy-tail distribution of flights. Furthermore, we find that humans perform a distance optimizing heuristic when planning trips over a set of destinations. This is another cause of the heavy tail flight distribution. We propose a bursty spot model (BSM) based on these observations, which is a simple human mobility model taking as input the degree of burstiness in visit point dispersion. We generate bursty hot spots by emulating the burstiness of visit points measured from real traces. We apply BSM to the evaluation of several DTN routing protocols and verify that our model generates heavy-tail flights. The emulation of hot spots reduces DTN routing performance substantially from popular random walk models– an order-of-magnitude less than LW and even 2 to 3 times less than RWP. This result indicates that the statistical features of hot spots and flights make significant impacts on the performance of routing in mobile networks and realistic representations of them are important.

8. REFERENCES

[1] Garmin GPSMAP 60CSx User's manual. http://www.garmin.com/products/gpsmap60csx/.

[2] F. Bai, N. Sadagopan, and A. Helmy. Important: a framework to systematically analyze the impact of mobility on performance of routing protocols for adhoc networks. In *Proc. of IEEE INFOCOM 2003*.

[3] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Nature*, 286:509–512, 1999.

[4] N. Biggs, E. LLoyd, and R. Wilson. *Graph Theory 1736-1936*. Clarendon Press, Oxford, 1976.

[5] V. Borrel, M. D. de Amorim, and S. Fdida. A preferential attachment gathering mobility model. *IEEE Communications Letters*, 9:900–902, 2005.

[6] P. L. M. de Maupertuis. ccord de differentes lois de la nature qui avaient jusqu'ici paru incompatibles. *Mem. As. Sc.*, page 417, 1744.

[7] J. Gaite. Zipfs law for fractal voids and a new void-finder. *The European Physical Journal B - Condensed Matter and Complex Systems*, 47:93–98, 2005.

[8] M. Grossglauser and D. N. C. Tse. Mobility increases the capacity of ad hoc wireless networks. *IEEE/ACM Trans. on Networking*, 10(4):477–486, 2002.

[9] T. Henderson, D. Kotz, and I. Abyzov. The changing usage of a mature campus-wide wireless network. In *Proc. of IEEE MobiCom 2004*.

[10] K. Herrmann. Modeling the sociological aspects of mobility in ad hoc networks. In *Proc. of MSWiM 2003*.

[11] X. Hong, M. Gerla, G. Pei, and C.-C. Chiang. A group mobility model for ad hoc wireless networks. In *Proc. of MSWiM 1999*.

[12] E. Hyytia, P. Lassila, and J. Virtamo. A markovian waypoint mobility model with application to hotspot modeling. In *Proc. of IEEE ICC 2006*.

[13] A. Jardosh, E. M. BeldingRoyer, K. C. Almeroth,

and S. Suri. Towards realistic mobility models for mobile ad hoc networks. In *Proc. of MobiCom 2003*.

[14] W. jen Hsu, K. Merchant, H. wei Shu, C. hsin Hsu, and A. Helmy. Preference-based mobility model and the case for congestion relief in wlans using ad hoc networks. In *Proc. of IEEE VTC 2004-Fall*.

[15] P. Juang, H. Oki, Y. Wang, M. Martonosi, L. S. Peh, and D. Rubenstein. Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with zebranet. In *ACM ASPLOS*, 2002.

[16] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In *Proc. of IEEE INFOCOM 2006*.

[17] S. Lim, C. Yu, and C. R. Das. Clustered mobility model for scale-free wireless networks. In *Proc. of IEEE LCN 2006*.

[18] A. Lindgren, A. Doria, and O. Schelen. Probabilistic routing in intermittently connected networks. *SIGMOBILE Mobile Computing and Communication Review*, 7(3), 2003.

[19] B. B. Mandelbrot. *Fractal Geometry of Nature*. W. H. Freeman anc Company, New York, 1977.

[20] M. McNett and G. M. Voelker. Access and mobility of wireless pda users. *Mobile Computing Communications Review*, 9:40–55, 2005.

[21] M. Musolesi, S. Hailes, and C. Mascolo. An ad hoc mobility model founded on social network theory. In *Proc. of MSWiM 2004*.

[22] C. Nilsson. Heuristics for the traveling salesman problem. In *Technical report, Linkoping University*, 2003.

[23] G. Ramos-Fernandez, J. L. Morales, O. Miramontes, G. Cocho, H. Larralde, and B. Ayala-Orozco. Levy walk patterns in the foraging movements of spider monkeys (ateles geoffroyi). *Behavioural Ecology and Sociobiology*, 55:223–230, 2004.

[24] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong. Human mobility patterns and their impact on delay tolerant networks. In *Proc. of HotNets-VI*, Atlanta, GA, November 2007.

[25] I. Rhee, M. Shin, S. Hong, I. Rhee, M. Shin, and S. Hong. On the levy-walk nature of human mobility. In *Proc. of IEEE INFOCOM 2008*, Pheonix, AZ, April 2008.

[26] D. J. Rosenkrantz, R. E. Stearns, and P. M. L. II. An analysis of several heuristics for the traveling salesman problem. *SIAM J. Comput.*, 6:563–581, 1977.

[27] T. Spyropoulos, K. Psounis, and C. Raghavendra. Efficient routing in intermittently connected mobile networks: The single-copy case. *IEEE/ACM Trans. on Networking*, 2007.

[28] G. M. Viswanathan, V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, and H. E. Stanley. Levy flights search patterns of wandering albatrosses. *Nature*, 381:413–415, 1996.