2009 International Conference on Computational Science and Engineering

Investigating Bimodal Clustering in Human Mobility

James P. Bagrow
Center for Complex Network Research
and Department of Physics, Northeastern University
Boston, MA 02115
Email: j.bagrow@neu.edu

Tal Koren
Center for Complex Network Research
and Department of Physics, Northeastern University
Boston, MA 02115
Email: t.koren@neu.edu

***Abstract*—We apply a simple clustering algorithm to a large dataset of cellular telecommunication records, reducing the complexity of mobile phone users' full trajectories and allowing for simple statistics to characterize their properties. For the case of two clusters, we quantify how clustered human mobility is, how much of a user's spatial dispersion is due to motion between clusters, and how spatially and temporally separated clusters are from one another.**

***Keywords*-human mobility; clustering; mobile phones**

I. Introduction

Recently, much effort has been devoted to understanding, mapping, and modeling large-scale human and animal trajectories [1]–[12]. Examples include models that describe agents searching a space for a target of unknown position [5], [6], [8]; mobility tracking in cellular environments [10], [13], [14]; human infectious disease dynamics and mobile phone viruses [9], [12], [15]–[17]; traffic transportation, and urban planning [2], and references therein. Understanding spatiotemporal patterns and characterizing human mobility and social interactions can now be achieved due to extensive and widespread use of wireless communication devices [10].

There is growing experimental evidence that the movement of many species, including humans, can be described by a class of non-trivial random walk models known as Lévy flights [1], [2], [18]. A Lévy flight can be considered a generalization of Brownian motion [18] and belongs to the class of scale-invarient, fractal random processes. Lévy flights are Markovian stochastic processes in which step lengths λ are drawn from a power law distribution:

$$\lambda(x) \simeq 1/|x|^{\alpha+1}, \tag{1}$$

where $0 < \alpha \leq 2$ is the Lévy exponent. This implies that the second moment of λ diverges and extremely long jumps are possible. For review, see [19].

Lévy statistics, somewhat controversially, have been found in the search behavior of many species including human hunter-gatherers [20]. Lévy flight-like movement patterns have been observed while tracking dollar bills [1], mobile phone users [2], and GPS trajectory traces obtained from taxicabs and volunteers in various outdoor settings [7], [21].

The mobile phone users studied in [2] reveal behavior similar to Lévy patterns, but individual trajectories show a high degree of temporal and spatial regularity. That investigation analyzed the trajectories of 10^5 anonymized phone users, randomly selected from more than six million subscribers. The unique mobility patterns found in [2] show a time-independent characteristic travel distance and high probability to frequently return to a few locations. Additionally, real user's mobility patterns may be approximated by Lévy flights but only up to a distance characterized by the radius of gyration r_g. This quantity represents the characteristic distance travelled by a user a observed up to time t, defined as

$$r_g^a(t) = \sqrt{\frac{1}{n_c^a(t)} \sum_{i=1}^{n_c^a} (\mathbf{r}_i^a - \mathbf{r}_{\mathrm{CM}}^a)^2}, \tag{2}$$

where $\mathbf{r}_i^a$ represents the $i = 1, \ldots, n_c^a(t)$ positions recorded for user a and $\mathbf{r}_{\mathrm{CM}}^a = 1/n_c^a(t) \sum_{i=1}^{n_c^a} \mathbf{r}_i^a$ is the center of mass of the trajectory [2]. Interestingly, it has been observed that the radius of gyration increases only logarithmically in time, which cannot be explained by traditional Lévy models; therefore, we must return to the data for further analysis.

A. Motivation and Open Questions

The work in [2] showed that human mobility patterns are well characterized by the radius of gyration r_g. This is a static measure, in the sense that the order of movement between locations is irrelevant, and can then characterize time-invariant properties of the trajectory. Mobile phone data only samples the actual underlying trajectory, however, with a user-driven, heterogeneous sampling rate [2], and this complicates the study of the user's real mobility. In the same way that r_g avoids these sampling problems by "integrating" over time, an appropriate spatial course-graining can provide a basic picture of the time-dependent, evolving characteristics of a subject's mobility pattern.

In this paper, we apply a simple clustering algorithm to the spatial locations of a user's trajectory. Finding clusters of frequently visited locations (such as home and work) and collapsing them to a single entity reduces the complexity of the full trajectory while allowing for simple statistics to capture properties relating to how users move between locations. Interesting questions include:

978-0-7695-3823-5/09 $26.00 © 2009 IEEE
DOI 10.1109/CSE.2009.283

1) How spatially separated are such clusters?
2) How often are clusters (re-)visited? How long do users dwell within clusters?
3) Are larger clusters (more recorded calls over time) more spatially dispersed (as quantified by the r_g of the cluster's elements) than smaller clusters? How do the r_g's of clusters relate to the total r_g?

The remainder of this paper is organized as follows. We first briefly discuss the dataset (Sec. I-B) and the clustering algorithm used to analyze it (Sec. I-C). We then introduce several important statistics, calculate them from the dataset at hand, and discuss their implications (Sec. II). A summary and discussion of future work follows (Sec. III).

B. Data Set

As in [2], [17] we analyze data from a European mobile phone carrier. The data contains the date and time of phone calls and text messages from 6 million anonymous users as well as the spatial location of the phone towers routing these communications. User locations within a tower's service area are not known. From this full dataset, we select a random subset of 60 000 users that make or receive at least one phone call during June – August 2007. The call history of each user was then used to reconstruct their trajectory of motion during that time period.

C. Clustering

Our analysis is based on k-means clustering which, in general, divides N-dimensional populations (or observations) into k distinct sets or clusters by minimizing the intra-cluster sum of squares w^2 of an appropriately defined distance metric. Using the notation in [22] (see also [23]):

$$w^2(S) = \sum_{i=1}^{k} \int_{S_i} |z - u_i|^2 \, \mathrm{d}p(z), \tag{3}$$

where p is the probability mass function for the observations $S = \{S_1, S_2, \ldots, S_k\}$, and u_i $(i = 1, \ldots, k)$ is the conditional mean of p over the set S_i. For this work, μ_i will be the center of mass of cluster i and we seek to find the locations of μ_i that minimize the square Euclidean distance between μ_i and that cluster's recorded call locations.

For simplicity, we assume $k = 2$ clusters throughout this work. This is a serious assumption, and identifying the correct number of clusters on a per-user basis remains important future work. However, we have found that the majority of users are well clustered with $k = 2$. To show this, we compute the mean *silhouette value* [24] for each user. Define $A(\mathbf{r}_i)$ as the average square (Euclidean) distance from $\mathbf{r}_i$ to all other points in the same cluster, and define $B(\mathbf{r}_i)$ as the average square distance from $\mathbf{r}_i$ to all points in the other cluster. The silhouette value $s(\mathbf{r}_i)$ for point $\mathbf{r}_i$ is

$$s(\mathbf{r}_i) \equiv \frac{B(\mathbf{r}_i) - A(\mathbf{r}_i)}{\max\{B(\mathbf{r}_i), A(\mathbf{r}_i)\}}, \tag{4}$$

and takes values between -1 and 1, with larger values indicating r_i is increasingly well separated from the other cluster. Taking the average $\langle s \rangle \equiv \langle s(\mathbf{r}_i) \rangle_i$ over all points then provides a single statistic measuring how well the whole data are clustered. Poor choices of k, for example, lead to smaller $\langle s \rangle$ [24]. We find that 91.8% of users have $\langle s \rangle > 0.8$ and 80.8% have $\langle s \rangle > 0.9$, indicating that the majority of users are well clustered with just $k = 2$.

II. Results

For each user, we apply the k-means algorithm to their trajectory, partitioning the call locations into $k = 2$ sets. The number of calls in clusters 1 and 2 for user a are $N_1(a)$ and $N_2(a)$, respectively (we identify $N_1(a) \geq N_2(a)$ such that cluster 1 is the primary cluster) and $N_T(a) = N_1(a) + N_2(a)$ is the total number of calls that user a makes during the sample period. The distribution $P(N)$ of the number of calls is shown in Fig. 1.

Using the spatial distribution of calls we compute each cluster's center of mass, $\mathbf{r}_{\mathrm{CM}}^{(1)}$ and $\mathbf{r}_{\mathrm{CM}}^{(2)}$, and radius of gyration, $r_g^{(1)}$ and $r_g^{(2)}$, as well as the total center of mass and radius of gyration for all points, $\mathbf{r}_{\mathrm{CM}}^{(T)}$, and $r_g^{(T)}$, respectively. The distributions of these quantities are shown in Fig. 2.

To quantify relationships between the two clusters, we compute the separation between their centers of mass, d_{CM}:

$$d_{\mathrm{CM}} = \left\| \mathbf{r}_{\mathrm{CM}}^{(1)} - \mathbf{r}_{\mathrm{CM}}^{(2)} \right\|, \tag{5}$$

where $\|\ldots\|$ is the Euclidean norm, and we also count how often a user "jumps" between the two clusters. A user who makes N_T calls will have $N_T - 1$ jumps between locations (including remaining at the current location). We define F_{CC} as the fraction of cross-cluster jumps, those that begin and end in different clusters. The distributions of F_{CC} and d_{CM} are shown in the insets of Figs. 1 and 2, respectively.

The number of calls per cluster indicates that users spend the majority of their time in one cluster and visit the other cluster more rarely. The fraction of cross-cluster jumps is small, $\langle F_{\mathrm{CC}} \rangle_{\mathrm{users}}$ is less than 0.1, (where $\langle \ldots \rangle_{\mathrm{users}}$ is an average over all sampled users), indicating that the primary cluster provides a stable location in which the user dwells. Likewise, d_{CM} is relatively large, $\langle d_{\mathrm{CM}} \rangle_{\mathrm{users}} = 157.8$ km, indicating that we are finding a semi-frequent but long-distance destination. It would be interesting to see temporal dependencies on the cluster's occupation probability: are users more likely to be in the secondary cluster on weekends, for example.

The cluster's radius of gyration summarizes how compact or dispersed user movement is within that cluster. We find that the larger cluster (in terms of the number of calls) tends to be slightly more spatially compact than the smaller cluster, $r_g^{(1)} < r_g^{(2)}$. Both $r_g^{(1)}$ and $r_g^{(2)}$ are much smaller than $r_g^{(T)}$, which is to be expected when d_{CM} is large. This means that much of the user's total radius of gyration is generated by

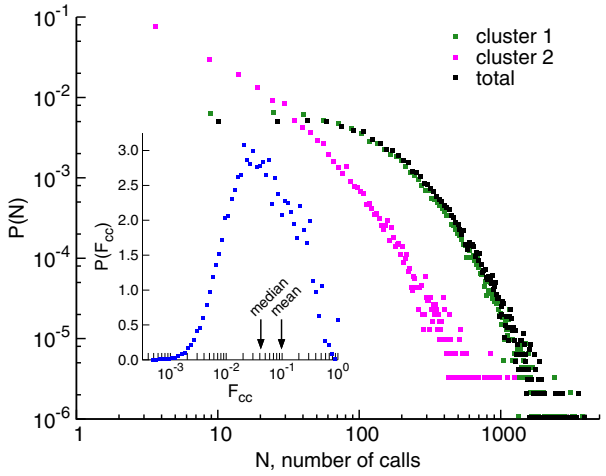

Figure 1: Temporal properties of clusters. Shown is the distribution $P(N)$ of the number of phone calls inside each cluster and the total number of calls, N_1, N_2, and $N_T = N_1 + N_2$, respectively. The majority of calls take place in cluster 1, but a non-negligible amount occur in cluster 2. (inset) The fraction of jumps $F_{\rm CC}$ from one cluster to another, quantifying how often users travel between their clusters. The primary cluster tends to contain the majority of calls and users tend to move between clusters somewhat rarely, $\langle F_{\rm CC} \rangle_{\rm users} = 0.098$, though some move much more frequently.

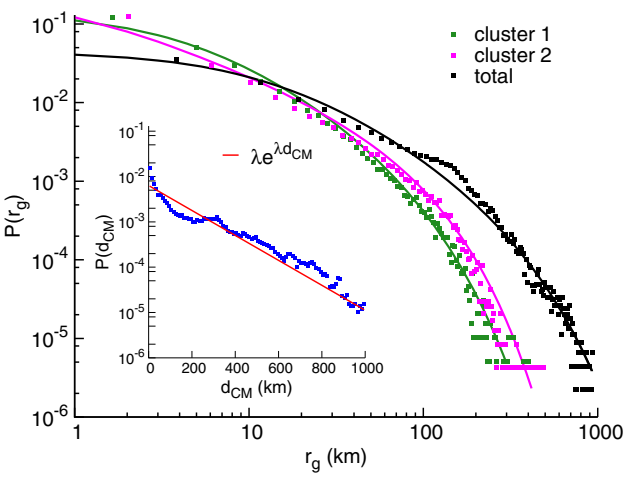

Figure 2: Spread and separation of trajectory clusters. Shown is the distribution $P(r_g)$ of the radii of gyration for both clusters and over all points, $r_g^{(1)}$, $r_g^{(2)}$, and $r_g^{(T)}$, respectively. The secondary cluster tends to be slightly more spatially dispersed than the primary cluster. Lines are truncated power laws of the form $\left(r_g + r_g^0\right)^{-\beta_r} e^{-r_g/\kappa}$, characterized by parameters $\Theta \equiv \left(r_g^0, \beta_r, \kappa\right)$. For the above curves, $\Theta_1 = (5.5, 1.5, 70)$, $\Theta_2 = (0.75, 0.9, 70)$, and $\Theta_T = (15, 1.4, 260)$, for cluster 1, cluster 2, and both, respectively. (inset) The distribution $P(d_{\rm CM})$ of distances between cluster's centers of mass, over all users. The straight line is an exponential distribution with mean $\lambda^{-1} = 157.8$ km, indicating that clusters are often well separated, but distances fall off rather quickly.

movement between two well-separated clusters, as opposed to homogeneous motion over a large space.

III. Conclusions and Future Work

We have applied a simple k-means clustering algorithm to a large sample of human trajectories generated from mobile phone records. Doing this characterizes how users move within their set of visited locations and we find that people tend to have one dense, primary cluster and one secondary, dispersed cluster. Course-graining a user's trajectory into clusters also quantifies how often users move between clusters and we find that users spend the majority of their time in the primary cluster but visit the secondary cluster semi-frequently. The clusters themselves tend to be well separated, indicating that the secondary cluster is a long-range destination, but the distribution of these distances over all users falls off exponentially quickly, compared to the total radius of gyration.

The most important avenue for future work involves relaxing the assumption of $k = 2$ clusters. While mean silhouette values have shown that the data are well characterized by two clusters, it remains to be seen if introducing more clusters improves the picture. Furthermore, since so much of a typical user's time is spent in the primary cluster, there remains the tantalizing possibility that further substructure is present within it. In other words, the secondary cluster may represent infrequent long-range trips while the primary cluster may represent the union of home and work clusters, or home and school. Information about important routines such as daily commuting may be contained within the primary cluster.

Acknowledgment

The authors would like to thank Marta González and Albert-László Barabási for fruitful discussions, and Pu Wang and Dashun Wang for providing data. This work was supported by the James S. McDonnell Foundation 21st Century Initiative in Studying Complex Systems; NSF within the Dynamic Data Driven Applications Systems (CNS-0540348), Information Technology Research (DMR-0426737), and IIS-0513650 programs; the Defense Threat Reduction Agency Award HDTRA1-08-1-0027; and the U.S. Office of Naval Research Award N00014-07-C. JPB gratefully acknowledges support from DTRA grant BRBAA07-J-2-0035.

REFERENCES

[1] D. Brockmann, L. Hunagel and T. Geisel. The scaling laws of human travel. *Nature* vol. 439 no. 7075 pp. 462-465 January 2006.

[2] M. C. González, C. A. Hidalgo and A.-L. Barabási. Understanding individual human mobility patterns. *Nature* vol. 453 no. 7196 pp. 779-782 June 2008.

[3] G. M. Viswanathan, S. V. Buldyrev, S. Havlin, M. G. E. da Luz, E. P. Raposo and H. E. Stanley. Optimizing the success of random searches. *Nature* vol. 401 no. 6756 pp. 911-914 October 1999.

[4] D. W. Sims *et al.* Scaling laws of marine predator search behaviour. *Nature* vol. 451 no. 7182 pp. 1098-1103 February 2008.

[5] O. Bénichou, M. Coppey, M. Moreau, P.-H. Suet and R. Voituriez. Optimal search strategies for hidden targets. *Phys. Rev. Lett* vol. 94 issue 19 pp. 198101 (4 pages) May, 2005.

[6] O. Bénichou, M. Coppey, M. Moreau and R. Voituriez. Intermittent search strategies: When losing time becomes efficient. *Europhys. Lett.* vol. 75, no. 2, pp. 349-354, July. 2006.

[7] I. Rhee, M. Shin, S. Hong, K. Lee and S. Chong. On the Lévy-walk nature of human mobility. In *Proceedings of INFOCOM*, Arizona, USA, April 2008.

[8] M. A Lomholt, T. Koren, R. Metzler and J. Klafter. Lévy strategies in intermittent search processes are advantageous. *Proc. Natl. Acad. Sci. USA* vol. 105, no. 32 pp. 11055-11059, August 2008.

[9] D. Brockmann and F. Theis. Money Circulation, trackable items, and the emergence of universal human mobility patterns. *Pervasive Computing* vol. 7 no. 4 pp.28-35, October 2008.

[10] F. Giannotti and D. Fedreschi (Eds.). *Mobility, Data Mining and Privacy*. Springer-Verlag Berlin Heidelberg, 2008.

[11] K. Lee, S. Hong, S. J. Kim, I. Rhee and S. Chong. SLAW: A mobility model for human walks. In *Proceedings of INFOCOM*, Rio de Janeiro, Brazil, April 2009.

[12] D. Brockmann. Human mobility and spatial disease dynamics. *in Review of Nonlinear Dynamics and Complexity*, H. G. Schuster (Ed.) Wiley-VCH, 2009.

[13] A. Bhattacharya and S. K. Das. LeZi-Update an information-theoretic framework for personal mobility tracking in PCS networks, *Wireless Network* vol. 8, no. 2/3 pp. 121-135, March-May 2002.

[14] M. Kim, D. Kotz and S. Kim. Extracting a mobility model from real user traces. In *Proceedings of INFOCOM*, Barcelona, Spain, April 2006.

[15] A. L. Lloyd and R. .M. May. How viruses spread among computers and people, *Science* vol. 292 no. 5520 pp. 1316-1317, May 2001.

[16] J. Kleinberg. The wireless epidemic. *Nature*, vol. 449 no. 7160 pp. 287-288, September 2007.

[17] P. Wang, M. C. González, C. A. Hidalgo and A.-L. Barabási. Understanding the spreading patterns of mobile phone viruses. *Science* vol. 324 no. 5930 pp. 1071-1076 May 2009.

[18] J. Klafter, M. F. Shlesinger and G. Zumofen. Beyond Brownian motion. *Physics Today* vol. 49 issue 2 pp. 33-39 February 1996.

[19] A. V. Chechkin, R. Metzler, V. Yu. Gonchar and J. Klafter. Introduction to the theory of Lévy flights. In *Anomalous Transport: Foundations and Applications* edited by R. Klages, I. M. Sokolov and G. Radons, Wiley-VCH, 2008.

[20] C. T. Brown, L. S. Liebovitch, and R. Glendon. Lévy flights in dobe ju/hoansi foraging patterns. *Human Ecology*, vol. 35, no. 1, pp. 129-138, Feb. 2007. [Online]. Available: http://dx.doi.org/10.1007/s10745-006-9083-4

[21] B. Jiang, J. Yin, and S. Zhao. Characterizing human mobility patterns over a large street network. [Online]. Available: http://arxiv.org/abs/0809.5001?context=physics. Sept. 2008.

[22] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press pp. 281-297 1967.

[23] G. Gan, C. Ma and J. Wu. *Data Clustering: Theory, Algorithms and Applications. ASA-SIAM Series on Statistics and Applied Probability*. SIAM, Philadelphia, ASA, Alexandria VA, 2007.

[24] P. J. Rousseeuw. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics* vol. 20 issue 1 pp. 53?65, November 1987.