Supplemental Information for SBP-BRIMS Submission

Fine-Scale Prediction of People's Home Location using Social Media Footprints

Hamdi Kavak[1][0000-0003-4307-2381], Daniele Vernon-Bido[1], and Jose J. Padilla[1]

[1] Virginia Modeling Analysis and Simulation Center, Suffolk VA 23435, USA
hkava001@odu.edu

1. Mobility Features

As mentioned, we calculate mobility features for all unique locations (based on cluster label ID) for each user. Here we use check-in (tweeting) ratios instead of actual number of check-ins to keep all features normalized between 0 and 1.

- **Check-in Ratio (CR):** Check-in Ratio is the measure of number of check-ins of a user at a location against total check-ins in all locations. A common assumption made about home location is that it is the most visited location [1-3] This certainly holds for continuously captured data [4] and is found to be an important feature for sparsely shared social media data [5].
- **Check-in Ratio during Midnight (MR):** Midnight check-in ratio looks at all midnight check-ins (12:00 AM - 07:00 AM) of a user and calculate the ratio of midnight check-ins per visited location. While the home is usually the last place before a person becomes stationary [6], social media users share their locations during midnight while they are outside their home as well [5].
- **Check-in Ratio of the Last Destination of a Day (EDR):** This feature captures the last destination of the day which is found to be important to predict home location [6]. We identify all last check-ins of days and calculate the ratio per location using tweets shared between 05:00 PM in the evening until 03:00 AM in the morning. We capture this last tweet with an intuitive grouping algorithm that merges first three hours of next day with previous day as shared in supplemental.
- **Check-in Ratio of the Last Destination of a Day with Inactive Midnight (EIDR):** This feature is very similar to the EDR feature but ignores days when a user shares tweets during midnight. This feature captures the assumption that the user ends the day at home and do not spend time at night outside [5]. We again use the grouping algorithm developed for the previous feature.
- **PageRank (PR, RPR):** PageRank [7] is a well-known graph measure to show the importance of nodes based on number of influential edges to them and is a decent predictor for home location Hu et al. We represent unique places as nodes and transition between places as edges based on consequent check-ins in the same day until 3 AM. We calculate both weighted PageRank and reverse weighted PageRank scores. Weights are based the number of transitions between nodes and reverse PageRank is captured by swapping source-destination pair.
- **Land Use Pattern (LU):** Land use patterns are designed by local governments to regulate the consumption of space by inhabitants. The assumption here is that, home location of a person has to be at a residential area. There are many codes describing different uses of the land. For the purpose of this study, we group land use of an area as *residential* and *non-residential*. We capture this as a feature according to official land-use maps from the case city.

- **Kilometer Distance from Most Checked-in Location (KM):** Previous studies [1-3] have shown the importance of most checked-in place when it comes to home location prediction. However, Hu et al. reports that most checked-in location, on its own, is not as effective for social media users. We argue that the distance to the most checked-in location might offer clues worth investigating because the home is the hub to transition other places [8].

2. Data description for Fig.2

Days	Start date	End Date	Total footprints used	Number of training/test instances generated	Number of users represented
7	2014-05-16	2014-05-23	26,132	2,886	383
14	2014-05-16	2014-05-30	56,762	5,589	470
21	2014-05-16	2014-06-06	90,422	8,681	535
30	2014-05-16	2014-06-15	128,322	11,815	681
90	2014-05-16	2014-08-14	419,950	31,669	850
180	2014-05-16	2014-11-12	759,163	53,795	1058
270	2014-05-16	2015-02-10	1,041,359	68,805	1,195

3. Group description for section 4

Condition	Number of users (G_n)	Condition	Number of users (G_r)
$0 \leq G_{n1} < 75$	316	$0 \leq G_{r1} < 0.6$	321
$75 \leq G_{n2} < 225$	331	$0.6 \leq G_{r2} < 1.4$	330
$225 \leq G_{n3} < 475$	298	$1.4 \leq G_{r3} < 2.75$	301
$475 \leq G_{n4}$	323	$2.75 \leq G_{r4}$	316

References

[1] E. Cho, S. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," *Proceedings of the 17th ACM SIGKDD ...,* pp. 1082-1090, 2011.

[2] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks," pp. 329-336, 2011.

[3] T. Pontes *et al.*, "Beware of what you share: Inferring home location in social networks," *Proceedings - 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012,* pp. 571-578, 2012.

[4] M. Lin, W.-J. Hsu, and Z. Q. Lee, "Predictability of individuals' mobility with high-resolution positioning data," *Ubicomp,* pp. 381-381, 2012.

[5] T. Hu, J. Luo, H. Kautz, and A. Sadilek, "Home Location Inference from Sparse and Noisy Data: Models and Applications," *Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015,* pp. 1382-1387, 2016.

[6] J. Krumm, "Inference Attacks on Location Tracks," *Pervasive Computing,* vol. 10, no. Pervasive, pp. 127-143, 2007.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab1999.

[8] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González, "Unravelling daily human mobility motifs," *Journal of the Royal Society, Interface / the Royal Society,* vol. 10, no. 84, pp. 20130246-20130246, 2013.