Analysis of Proximity Networks from Multiple Mobile Sensor Data

Dae Wook Kim and Mehmet M. Dalkilic

***Abstract*—In this paper, we report our data-driven investigation on measuring proximity networks derived from multiple mobile sensor data: Bluetooth, GPS, and WLAN coordinates. By exploiting the dataset from NOKIA Mobile Data Challenge 2012, we propose our methods to generate sensor data based proximity networks between mobile users and evaluate these networks with a call log network as ground truth. Our results can be useful to understand that Bluetooth, GPS, and WLAN AP are of significant sensor data to quantify the pattern and strength of proximate relation between mobile users.**

***Index Terms*—Bluetooth, GPS, proximity network, WLAN AP.**

I. Introduction

The most fundamental challenges in computational social science [1] are unraveling social interactions and understanding social contexts. Mobile phone data allows us to detect social proximity by providing rich information such as call logs, mobility, and other sensor data. Specifically, Bluetooth (BT) devices have become a well-recognized and efficient sensor for constructing a proximity network. Accordingly, a large number of BT-based methods have been developed and applied to various contextual data over the past decade. Despite the considerable efforts to date, discovering the proximity patterns between people over time by BT-based methods is still a challenge [2],[3].Hence, to improve the detection of physical contacts between people, one must explore the relationship between BT data and additional contextual information from mobile phone data because GPS and WLAN AP coordinates capture rich aspects of a person's daily life.

In this paper, we present simple framework to discover proximity networks on Bluetooth, GPS, and WLAN coordinates and measure their accuracy based on call log network developed by ground truth. Our study focuses on the following contributions: (1) we introduce simple methods to construct a proximity network with Bluetooth and with geographic coordinates; (2) we evaluate our proximity networks for discovering proximate relation of 15 users identified by a call log network. The rest of the paper is organized as follows: Section 2 explains the raw dataset and call log network we analyzed. Section 3 introduces the method used to detect a Bluetooth proximity network. Section 4 presents a proximity network with geographic coordinates of GPS and WLAN AP. Section 5 evaluates the performance of Bluetooth, GPS, and WLAN AP proximity networks on identifying cell-phone communication. We conclude in Section 6 with some final remarks.

Manuscript received September 24, 2012; revised November 23, 2012.

Dae Wook Kim and Mehmet M. Dalkilic are with the School of Informatics and Computing, Indiana University, Bloomington, IN47405 U.S.A. (e-mail: daewkim@ cs.indiana.edu).

II. Data and Preparation

A. Large-Scale Mobile Sensor Data

Large-scale mobile sensor data used for our study were collected by Nokia mobile data collection campaign 2012 [4].We choose sensor data relevant to estimating proximity among users: Bluetooth, GPS, and WLAN AP. Among 38 users in the dataset, we focus on 15 users because we do not have call logs between other 23 users. We analyzed GPS and WLAN AP distribution on all the users. Fig. 1 captures majority in mobile users is frequently discovered between longitudes range from 6.5 to 7.0 and latitudes range from 46.4 and 46.6.

B. Call log Network as Ground Truth

We create a directed call log graph, which will be used as ground truth for the existing social interaction. The call log network is constructed by the anonymized call logs of 15 users since we can identify only 15 users from our raw data.The call log network consists of three different subgraphs: Blue, Yellow, and Red (See Fig. 2). Each node is labeled with user ID and the size of a node represents the total number of incoming or outgoing calls. Each edge is the bi-direction of a calling between users. We use this calllog network to evaluate the accuracy of proximity networks from Bluetooth, GPS, and WLAN AP.

III. Proximity Network With Bluetooth

A. Bluetooth Frequency of Mobile Users

To construct a proximity network with Bluetooth, we captured the Bluetooth frequency list observed for each user using anonymized Bluetooth mac addresses of 11 identified user IDs. In simple terms, we counted the number of occurrences of Bluetooth mac addresses observed for each user. As shown Fig. 3, a mobile user 51 contains Bluetooth frequencies from identified 8 mobile users: {2, 7, 17, 23, 56, 75, 89, 139}, with the mobile user 89 having the highest BT frequency (609) of the group.

B. Construct Bluetooth Proximity Network

We construct the Bluetooth proximity network using the anonymized Bluetooth mac addresses of each user. Based on the Bluetooth frequency list for each user, we create an edge from A to B when A's device sees B's device. Fig. 4 shows a directed network that represents the proximity of mobile users by Bluetooth device; the edge represents the proximity

between users, whereas the node size depends on the frequency of co-occurring Bluetooth mac addresses for each user. We detected that our Bluetooth proximity network was deployed at three distinct proximity subgraphs corresponding to the call log network. For instance, a subgraph connected with the following set of nodes: {17, 50, 56} of the Bluetooth proximity network is similar as a yellow subgraph of the call log network.

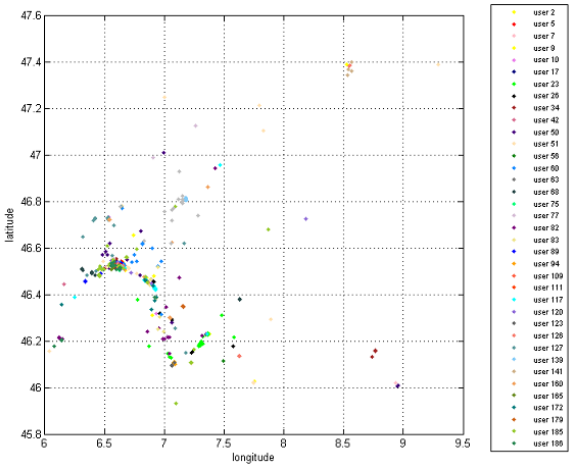

Fig. 1. WLAN AP distribution of 38 users.

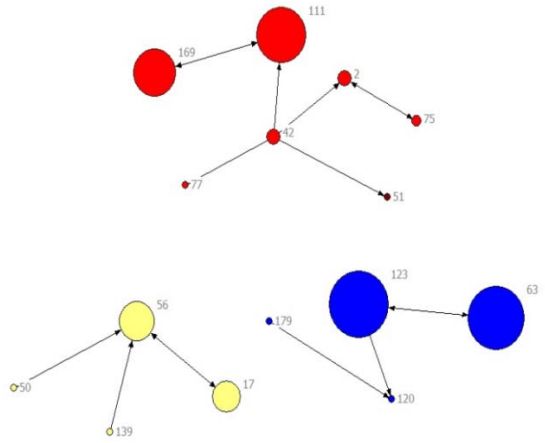

Fig. 2. Call log network from the 15 users.

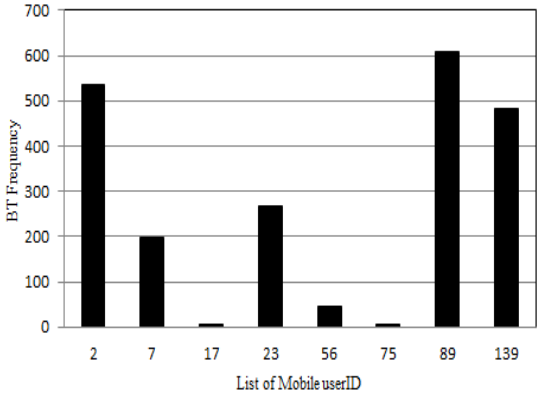

Fig. 3. Bluetooth frequency list of mobile user 51.

IV. Proximity Network Using Geographic Coordinates

A. Method

In this section, we briefly describe our method to construct a proximity network using geographical coordinates of GPS and WLAN AP respectively. By extracting the pair wise proximity frequencies between users, we consider the only accumulated proximity network. There are four main steps to discover GPS or WLAN AP proximity between users at a given period.

- *Data representation:* As a data preprocessing step, we extract a triplet representation: {unix time, longitude, latitude} from the raw GPS (or WLAN AP) data between users. Using (1), we calculate the minimum unix time interval (*t*) between two users:

$$t = \min\{|a - b| : a \in U_i,\ b \in U_j\} \quad (1)$$

where *a* and *b* are unix time from user *i*,U_i and user *j*,U_j respectively. In order to compute GPS distance between two users, the feature set (*F*) derived by *t* is expressed as follows:

$$F = [long_i lat_i \quad long_j \quad lat_j] \quad (2)$$

where *long* is a longitude and *lat* is a latitude of user *i* and *j* respectively.

- *GPS distance calculation:* We use the Carlson model [5][6] to calculate the GPS distance between two *points* using GPS coordinates. Fig. 5 depicts the GPS distance distribution of two users 2 and 51 based on our feature set *F* and Fig. 6 represents frequencies of GPS distance in specific intervals of GPS distance between user 2 and 51.
- *Proximity determination:* We set a user-defined threshold (1 meter) to determine useful GPS distance between users because we assume a physically contact distance is less than 1 meter. If the GPS distance between users is below the threshold, we define each user in their proximity with whom they could be personally known (See Fig. 7).
- *Construct proximity network:* Based on the information of users' proximity from the previous step, we build an accumulated proximity network in which *each* node represents a user ID and each edge is proximity between users.

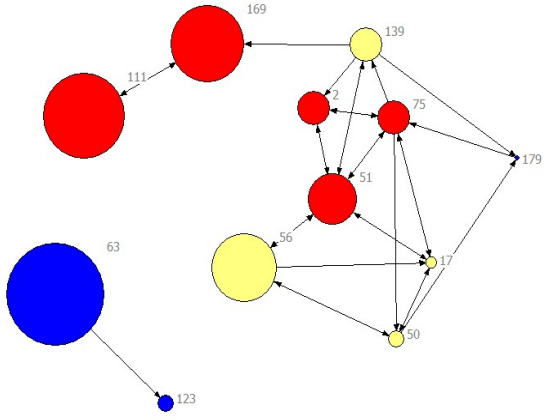

Fig. 4. Bluetooth proximity network for 11 identified mobile users.

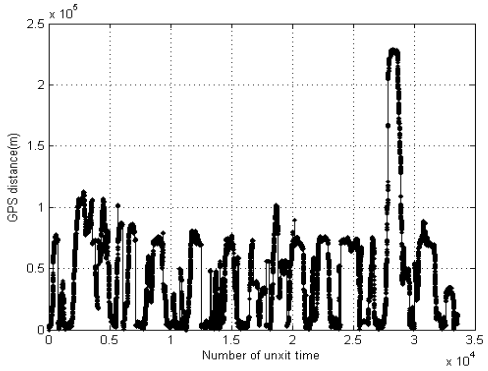

Fig. 5. GPS distance distribution between user 2 and 51.

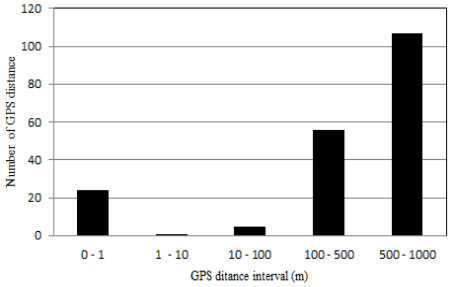

Fig. 6. Number of GPS distance on some GPS distance intervals between user 2 and user 51. Note that the number of their physically contact distance is 24.

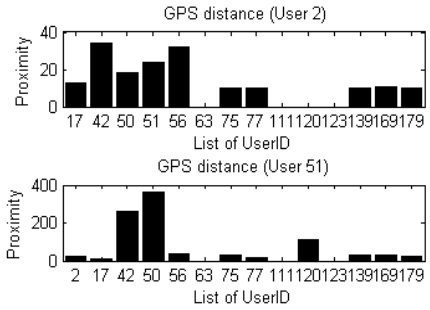

Fig. 7. Proximity distribution of user 2 a

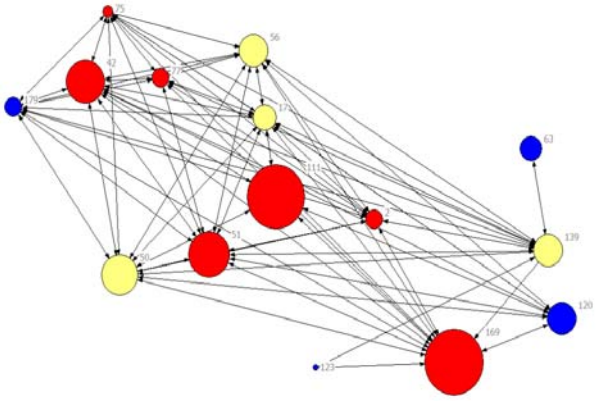

Fig. 8. Proximity network using GPS coordinates.

B. GPS or WLAN AP Proximity Network

We exploit GPS and WLAN AP coordinates from raw sensor data and built two types of proximity networks: a GPS based proximity network and a WLAN AP based proximity network. Fig. 8 represents the proximity network that is generated using only GPS coordinates and we can identify proximate relation between 15 users. We can detect the proximity between all the users except interactions between users 63, 120 and 179 (Blue). Fig. 9 shows the proximity network using WLAN AP coordinates and it discovers proximity of 13 users.

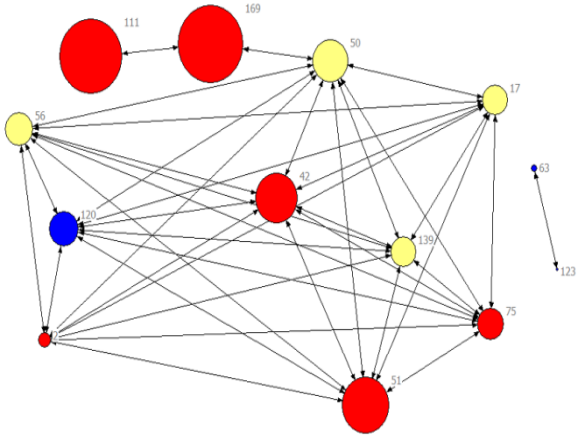

Fig. 9. Proximity network using WLAN AP coordinates.
nd user 51 respectively.

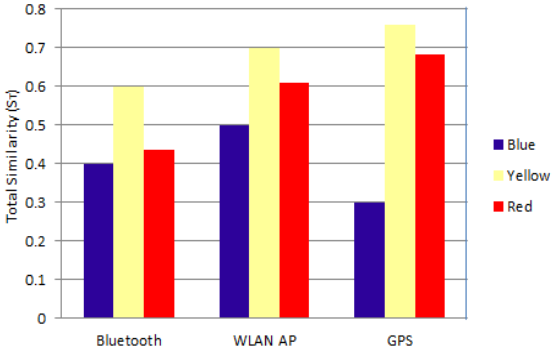

Fig. 10. Total similarity score S_T of proximity networks.

V. Evaluation

We develop a new measure to assess the topological similarity [7] of three subgraphs in our generated networks. To compare Bluetooth, GPS, and WLAN AP proximity networks with a call log network, we use the linear combination of edge-based similarity score and node-based similarity score [8]:

$$S_T = \sigma S_e + (1-\sigma) S_n (\sigma \in [0,1]) \quad (3)$$

where S_T is the total proximity similarity composed of S_e and S_n. S_e is the edge-based similarity and S_n is the node-based similarity between each of three proximity networks and the call log network. We calculate proximity similarities (i.e., $S_e$andS_n) using Jaccard's coefficient of similarity [9]. For the edge-based similarityS_e, we use the following equation to compute a similarity score (S) between each subgraph (i.e., blue, yellow, red) of a proximity network and the call log network respectively:

$$S = \frac{r}{p+q+r} \quad (4)$$

where p is the number of edges which are in a proximity network but not in a call log network, q is the number of edges which are in a call log network but not in a proximity network, and r is the number of edges which are in both. Similarity, the node-based similarity S_n is calculated based on the number of co-occurring nodes in each of the proximity networks and in the call log network. Fig. 10 shows the

overall similarity measurement from three different proximity networks in terms of (3).

When we set the value of σ as 0.6, we can achieve the highest performance based on our proposed measure. From the above result, we can capture that GPS proximity network and WLAN AP proximity network show the highest similarity in the yellow subgraph. GPS proximity network represents better performance than the other two proximity networks for a yellow subgraph and a red subgraph, whereas Bluetooth proximity network has the lowest similarity in the blue and red subgraphs. All the proximity networks show the poor performance of the blue subgraph.We also analyze the proximity networks and a call log network to extract the strongest close use pair between users by the number of the pair wise proximity frequencies between nodes from each of the proximity networks, or by the size of nodes from the call log network. We measure the strength of closeness between nodes by a point wise mutual information (PMI) [10] defined as:

$$PMI(x;y) = log\frac{p(x,y)}{p(x)p(y)} \quad (5)$$

where

$$p(x,y) = \frac{min\,|Size\ of\ a\ node\ with\ x, Size\ of\ a\ node\ with\ y|}{Size\ of\ total\ nodes},$$

$$p(x) = \frac{Size\ of\ a\ node\ with\ x}{Size\ of\ total\ nodes},$$

$$p(y) = \frac{Size\ of\ a\ node\ with\ y}{Size\ of\ total\ nodes}.$$

The PMI indicates the probability that two users are more familiar against the probability that they are less familiar. We minimize to calculate the frequencies of co-occurring node *x* and *y*. As shown in Table I, we discovered a strong correlation between the pair wise proximity frequencies influenced by Bluetooth, WLAN AP or GPS and the closeness of user's social contacts by call logs. Note that the user pairs who have the highest relationship in a call log network are the same as the user pairs in a Bluetooth proximate network.

TABLE I: USER PAIRS WITH THE STRONGEST CLOSENESS

	Blue	Yellow	Red
Call log	(63, 123)	(17, 56)	(111, 169)
Bluetooth	(63, 123)	(17, 56)	(111, 169)
WLAN AP	(63, 123)	(50, 56)	(111, 169)
GPS	None	(50, 56)	(111, 169)

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel approach for generating accumulated proximity networks from mobile sensor data and a call log network as ground truth in order to investigate the proximity of mobile users using NOKIA Lausanne data collection. The validation methods which apply for our analyses confirm that Bluetooth, WLAN AP, and GPS are critical sensor data todis cover the proximate relation between mobile users.

In future work, the proposed framework can be extended to develop a classification learner with feature sets derived from the combination of multiple sensor data such as Bluetooth, WLAN AP, and GPS. We can potentially discover the dynamics of the proximity networks by means of a classification learner. By comparing the proximate patterns of people between a call log network and our predicted proximity networks for a short duration (e.g., one consecutive hour in a day), we plan to show the accuracy of our classification learner-based proximity networks.

Another direction is that it would be of interest to develop a novel data-driven model to improve the reliability of our proposed detection mechanism by Bluetooth. This will allow a reduction of the limits(i.e., missing edges from Bluetooth device, scale of noise, and so forth) of current proximity network-generated model, and to devise a more realistic modeling scheme [11]. This is also part of our future work.

ACKNOWLEGMENTS

The authors thank the NOKIA Mobile Data Challenge (MDC) for providing the raw MDC dataset and also thank the anonymous reviewers for their helpful suggestions.

REFERENCES

[1] D. Lazer, *et al*., "Computational Social Science," *Science*, vol. 323, no. 5915, pp. 721–723, Feb. 2009.
[2] T. Do and D. G. Perez, "Human Interaction Discovery in Smartphone Proximity Networks," *Personal and Ubiquitous Computing*, published online Dec. 2011.
[3] T. Do and D. G. Perez, "Contextual grouping: discovering real-life interaction types from longitudinal Bluetooth data," in *Proc. of 12th IEEE Int. Conf. on Mobile Data Management*, Lulea, 2011, pp. 256-265.
[4] J. K. Laurila, *et al*., "The mobile data challenge: Big data formobile computing research," in *Proc. of Mobile Data Challenge by Nokia Workshop, in conjunction with Int. Conf. on 10th Pervasive Computing*, Newcastle, June 2012.
[5] C. G. Carlson, "What do latitude and longitude, readings from a gps receiver mean?" Technical report, Dept. Plant. Sci., South Dakota State Univ., Brookings, SD, 1999.
[6] C. G. Carlson and D. E. Clay, "The Earth Model – Calculating Field Size and Distances between Points using GPS Coordinates," *Site-Specific Management Guidelinesseries-11*, Potash and Phosphate Institute (PPI), 1999.
[7] S. Erten, G. Bebek, and M. Koyuturk, "Disease gene prioritization based on topological similarity in protein-protein interaction networks," in *Proc. 15th Annual Int. Conf. on Research in Computational Molecular Biology*, Vancouver, 2011, pp. 54-68.
[8] X. M. Jiang, G. R. Xue, W. G. Song, H. J. Zeng, Z. Chen, and W. Y. Ma, "Exploiting Page Rank at Different Block Level," in *Proc. of 5th Int. Conf. on Web Information Systems Engineering*, Brisbane, 2004, pp. 241–252.
[9] P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bul. Soc. Vaudoise Sci. Nat.*, vol. 44, pp. 223-270, 1908.
[10] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic, "Recipe recommendation using ingredient networks," in *Proc. 4th Int. ACM Web Science Conf.*, Evanston, 2012, pp. 447-456.
[11] E. Yoneki, "The Importance of Data Collection for Modelling Contact Networks," in *Proc. 12th IEEE Int. Conf. on Computational Science and Engineering*, Vancouver, 2009, pp. 940–943.

Dae Wook Kim has received his B.S. in computer science and engineering from Michigan State University, East lansing, Michigan, U.S.A. in 2003 and his M.S. in computer science and information from Syracuse University, Syracuse, New York, U.S.A. in 2007. Currently, he is pursuing his Ph.D. in computer science from the School of Informatics and Computing from Indiana University, Bloomington, Indiana, U.S.A. His research interests include data mining, ontology, database theory and system.

Mehmet M. Dalkilic has received his B.A. in chemistry with honors in 1998and his M.S. in computer science in 1996 and has done his Ph.D. in computer science in 2000 from Indiana University, Bloomington, Indiana, U.S.A. Currently, he is working as an ASSOCIATE PROFESSOR in the School of Informatics and Computing, Bloomington, Indiana, U.S.A. His research interests include data mining, ontology, database theory and system, bioinformatics. Dr. Mehmet M. Dalkilic is a member of ACM, IEEE, VLDB, PSB and ISMB.