Temporal Analysis of Online Social Graph by Home Location

Shiori Hironaka **Mitsuo Yoshida** **Kyoji Umemura**
Toyohashi University of Technology
Aichi, Japan
s143369@edu.tut.ac.jp, yoshida@cs.tut.ac.jp, umemura@tut.jp

ABSTRACT
An online social graph which represents relationships between users is used for many purposes such as home location estimation. However, the online social graph changes over time because the user's environment changes (e.g., house-moving). We tackle temporal analysis of online social graphs by answering the following question: which social graph of certain periods shows the best performance for network-based home location estimation on Twitter? We obtain that the estimation performance achieves the best using the social graph after about half a year. This result indicates that changes in social graphs due to user's environmental changes converge after about half a year.

ACM Classification Keywords
J.4. Computer Applications: Social and Behavioral Sciences

Author Keywords
Twitter; social graph; home location estimation

INTRODUCTION
People live constructing relationships and interacting with each other. An online social graph captures a realistic social graph constructed from such relationships [3]. Therefore, an online social graph is used for many purposes, especially to estimate user attributes such as home locations [1, 4]. The home location estimation methods using online social graphs are called *network-based home location estimation* methods.

The online social graph changes over time because the user's environment changes (e.g., house-moving). We have to update online social graph data [6], and the estimation performance of network-based home location estimation may change depending on when we collect the online social graph data. McGee et al. [5] indicated that the geographic distance changes by the relationship between users, which constitutes an online social graph. Is a newer online social graph used for home location estimation better?

WII'18, March 11, 2018, Tokyo, Japan

In this paper, we tackle temporal analysis of online social graphs by answering the following question: which social graph of certain periods shows the best performance for network-based home location estimation on Twitter? We obtain that the estimation performance achieves the best using the social graph after about half a year. This result indicates that changes in social graphs due to user's environmental changes converge after about half a year.

NETWORK-BASED HOME LOCATION ESTIMATION
A network-based home location estimation method is the home location estimation method using a social graph, which is created with a node as a user and an edge as a relationship between users. The based assumption is that a user is located geographically close to friends on the social graph. We use network-based home location estimation to determine how well the social graph reflects home locations (whether the social graph and home location data represent the state at the same time).

In this paper, we use the method of Davis Jr. et al. [2] as a popular network-based home location estimation. This method selects the most frequent location among the locations of the user's friends as the estimated location. The method is represented as follows:

$$S_u = \underset{l \in \{l_n | n \in N_u \cap L\}}{\arg\max^*} |\{v | v \in N_u \cap L, l = l_v\}|$$

$$\mathrm{Infer}(u) = \underset{l \in S_u}{\arg\max} |\{n | n \in L, l = l_n\}|$$

where L is a set of learning data (nodes), N_u is a set of adjacent nodes of node u, l_u is a correct label (home location) of node u, and $\arg\max^*$ is defined that returns a set of the equivalent. The processing that the maximum value of the number of friends' locations is the equivalent is not clear in the paper [2]. In this paper, we then prepare a set S_u of l which takes the maximum value to select the most frequent area in the learning data set.

DATA
We need home location data and social graph data for home location estimation. In this section, we describe how to make the data.

We define that a user's home location is the most frequent location posted with geo-tagged tweets by the user. Actually, we decided that a home location is an administrative area like a city, the same way as Davis Jr. et al. [2]. We aggregate the

locations of the geo-tagged tweets for each area, and select the most frequent city as a user's home location.

We collected geo-tagged tweets posted in Japan from January 2014 to December 2016 using the Twitter Streaming API. We used only the tweets which have "place" field and its "place_type" is "poi" or "city". We regard a city in Japan including the centroid of the bounding box of the "place" as a location of the geo-tagged tweet. We assigned a home location to a user who posts geo-tagged tweets at least five times each year in a certain area in Japan. As a result, we assigned a home location to 634,789 users in 2014, 851,675 users in 2015, and 828,929 users in 2016.

In this paper, we use a social graph based on a mutual *following* relationship on Twitter. Our social graph is a simple undirected graph. We collected *following* relationships every month from July 2015 to July 2017 among users who were assigned a home location in 2014. We excluded users whose following relationships could not be collected one or more times due to account deletion or becoming a private account. We collected following relationships and created 25 monthly social graphs.

Finally, we use 76,730 users for analysis, who can be assigned a home location for the three years continuously and whose following relationships can be collected for those three years.

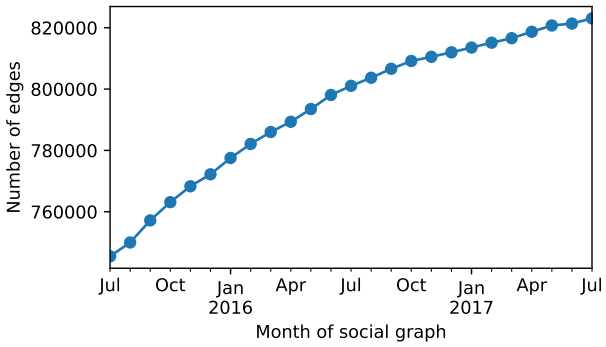
ANALYSIS

Temporal Analysis of Home Location and Social Graph

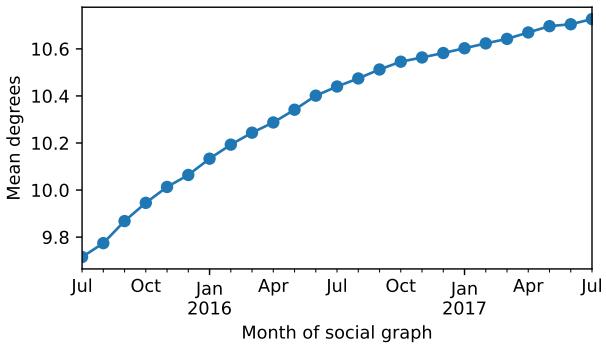
In this section, we report that social graphs and home locations change over time.

Firstly, we report the changes of the user's home locations from 2014 to 2016. In the 76,730 target users, the home locations of 39,814 users did not change for the three years, the home locations of 22,477 users changed once, and the home locations of 14,439 users changed twice. In total, the home locations of 48% of users changed at least once.

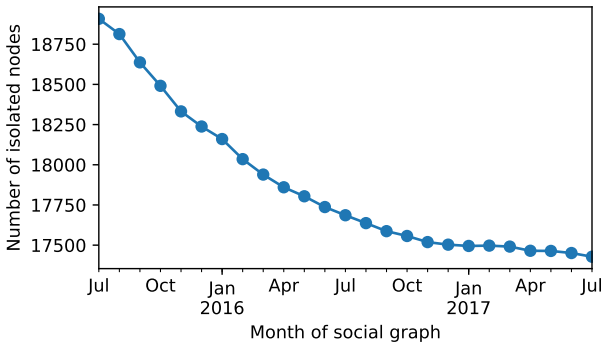
Secondly, we report the changes of the social graph from July 2015 to July 2017. The changes of the social graph properties are shown in Figure 1. Figure 1(a) shows that the number of edges between 76,730 users increases over time. Figure 1(b) shows the average number of degrees, which is used as an estimation clue, increases from 9.7 to 10.7 for two years. Figure 1(c) shows that the number of isolated nodes, which have no edges, decreases over time. They therefore show that we can use more relationships (edges) to estimate home locations in after months and years.



(a) number of edges



(b) average number of degrees



(c) number of isolated nodes

Figure 1: Size of social graphs: the number of edges increases and the number of isolated nodes decreases over time.

Comparison of Social Graph Collected Month

The network-based home location estimation method uses the combination of home locations and a social graph. In this paper, the home location is created from a certain period of geo-tagged tweets, and the social graph is a monthly snapshot. We investigate that the estimation performance combining old home location data and old social graph data, old home location data and new social graph data, and new home location data and new social graph data. We did not use old home location data to estimate new home location in this analysis. That is, the home location data for learning and tests are from the same period of data. In this analysis, since it is considered that the network-based home location estimation is making a good guess using the social graph and friends' locations, we find the most adequate period of the social graph for determining home locations.

We conduct a home location estimation combining three-years of home locations and 25 monthly snapshots of a social graph for two years. Since we are interested only in whether a user's home location can be estimated correctly, the performance is measured by precision, recall, and F1 with leave-one-out cross-validation.

The results of the estimations are shown in Figure 2. When we use the home locations of 2014, August 2015 achieved the highest F1. In the home locations of 2015, F1 increases to June 2016, and decreases after that. In the home locations of 2016, F1 becomes higher to June 2017. These results show the highest values of precision, recall, and F1 are achieved after about half a year from the end of the year when home locations were assigned. This result indicates that changes in social graphs due to user's environmental changes converge after about half a year.

The results also show poor performance when using the home locations of 2016. We conjecture that this cause is due to a change of default function for location acquisition by Twitter in April 2015. The default function has been changed from the accurate GPS coordinates to the user's self-chosen place[1]. As a result, the quality of "place" used to assign home location has declined.

Analysis of Users Who Have Changed Home Location

We assume that the performance changes shown in Figure 2 are caused by home location changes. We evaluate the performance of home location estimation by splitting into two user groups: users who have changed their home location at least one time within the three years and users who have never changed their home location within those three years. We did not distinguish between the home location changes that occurred once or twice. We show just evaluation result of F1 because precision and recall show the same trend as F1.

Figure 3 shows the evaluation result of F1 by splitting into the above two user groups. The estimation results of users who have changed their location have large differences over time. In contrast, the estimation results of users who have not changed home location have few differences. When we consider the reason for location change is house-moving, it seems home location changes cause social graph changes.

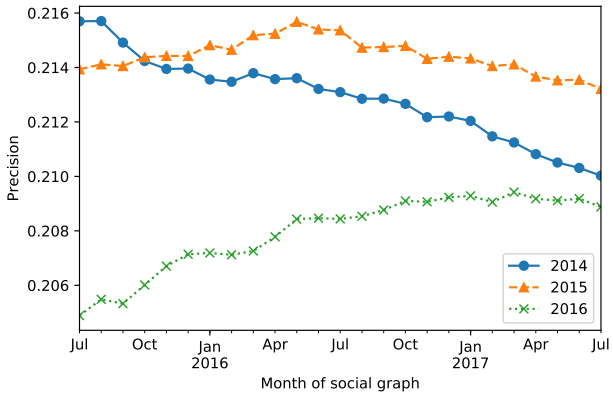
In addition, in contrast to the averages of F1 of users whose location are stable are 0.241, 0.243, and 0.233 in 2014, 2015, and 2016 respectively the averages of F1 of users whose location are changed are respectively, 0.127, 0.127, and 0.126. This result indicates that users who have changed home location are hard to estimate using this network-based home location estimation method.

Analysis of Users Who Have Changed Friends

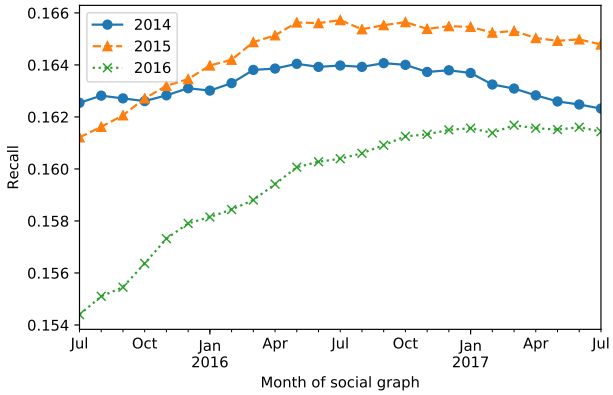
It is considered that users whose social graph changes make new friends actively and we suppose that the change of social graph converges quickly, specifically before about half a year. We evaluate the performance of home location estimation by splitting into two user groups: users who have changed friends (adjacent nodes) within the three years and users who have not changed friends.

To check the change of friends, we compare the number of friends (degree) of each user on the social graph July 2015 and July 2017. This metric watches only mutual following friends. In this case, a user just a *followee* or *follower* is not a friend.

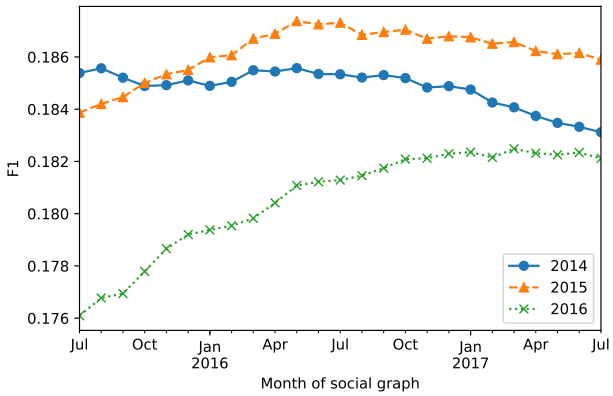
[1]The number of tweets which have "coordinates" field sharply decreased on April 28, 2015.



(a) Precision



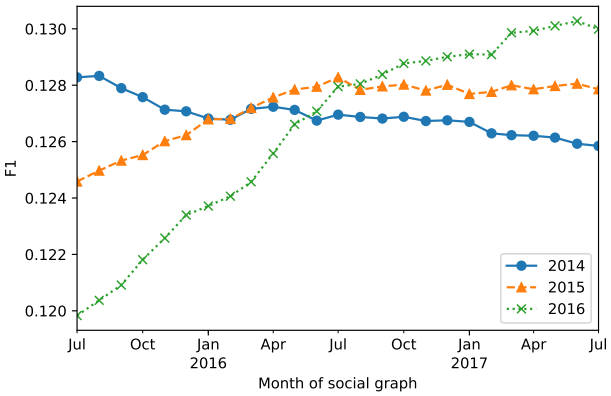
(b) Recall



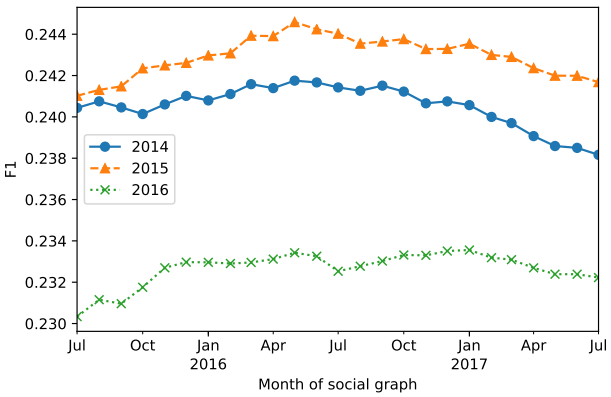
(c) F1

Figure 2: Estimation performance combining home location of each year and social graphs of each month. The highest F1 value is achieved after about half a year from the end of the year when home locations were assigned.

The number of users who have changed the number of friends is 44,228, and the number of users who have not changed the number of friends is 32,502.

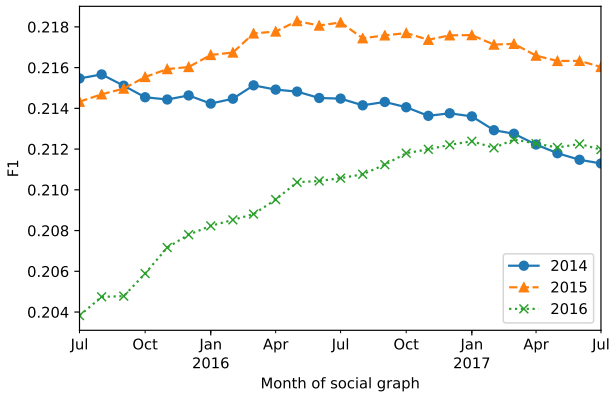


(a) The users having moved location ($n = 36916$)

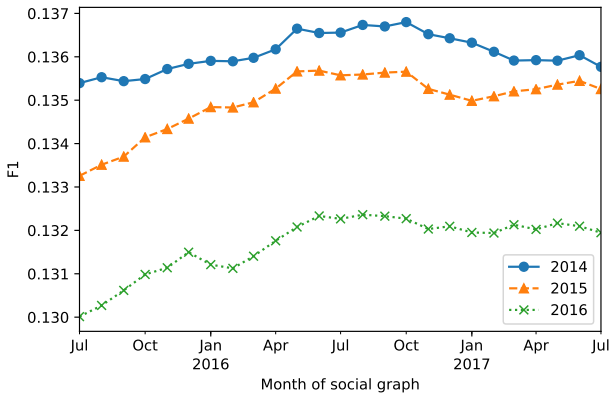


(b) The users having stable location ($n = 39814$)

Figure 3: Estimation performance comparison with stable location and moved location.



(a) The results of users who have changed friends ($n = 44228$)



(b) The results of users who have not changed friends ($n = 32502$)

Figure 4: Comparison of social graph changes and estimation performance.

The evaluation result is shown in Figure 4. The estimation results of users who have changed friends have large differences. In contrast, the estimation results of users who have not changed friends have few differences. We surmise the F1 of users making friends actively finish to change before about half a year, but we cannot observe differences in convergence speed compared with Figure 2. The average F1 of users who have changed friends is 0.214, 0.217, and 0.210 in 2014, 2015, and 2016 respectively, and the average F1 of users who have not changed friends is 0.136, 0.135, and 0.132 in 2014, 2015, and 2016 respectively. The users who have not changed friends have a lower F1.

DISCUSSION

Our research question was "Which social graph of certain periods shows the best performance for network-based home location estimation on Twitter?". We obtained the result that it is after about half a year from the end of the year when home locations were assigned. Our home location assigned method uses a year's worth of geo-tagged tweets. Thus, our result showing a peak after about half a year means a wide variance between about half a year to a year. The revealing of more detailed timing is a future work.

In the experiment, we reveal that social graph changes for about half a year to a year after a home location changes. When a user's home location changes, the home location data of a user changes when the majority of tweets during that year change to a new place. When the social graph data of a user changes when the majority of friends become new friends, the estimation result changes. Our results show that social graph changes are slower than home location changes. It is considered that the social graph changes significantly when the home location is changed.

CONCLUSION

We tackle temporal analysis of online social graphs by answering the following question: which social graph of certain periods shows the best performance for network-based home location estimation on Twitter? We collected monthly snapshots of a social graph for two years and user's home locations for three years. Using the data, we conduct home location estimation. We have obtained that the F1 achieved the highest performance after about half a year from the end of the year when home locations were assigned. In addition, we have found that these results can be seen in only users who have changed their home location at least once in three years.

REFERENCES

1. Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity. In *Proceedings of the 19th International Conference on World Wide Web*. 61–70.
2. Clodoveu A. Davis Jr., Gisele L. Pappa, Diogo Rennó Rocha de Oliveira, and Filipe de L. Arcanjo. 2011. Inferring the Location of Twitter Messages Based on User Relationships. *Transactions in GIS* 15, 6 (2011), 735–751.
3. Ravi Kumar, Jasmine Novak, and Andrew Tomkins. 2010. Structure and Evolution of Online Social Networks. In *Link Mining: Models, Algorithms, and Applications*. 337–357.
4. Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. 2012. Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations. In *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining*. 1023–1031.
5. Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. 2013. Location Prediction in Social Media Based on Tie Strength. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*. 459–468.
6. Norases Vesdapunt and Hector Garcia-Molina. 2016. Updating an Existing Social Graph Snapshot via a Limited API. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 1693–1702.