Information Sciences 422 (2018) 188–203

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Hidden community identification in location-based social network via probabilistic venue sequences

Hui Li[a,*], Ke Deng[b], Jiangtao Cui[c], Zhenhua Dong[d], Jianfeng Ma[a], Jianbin Huang[e]

[a] School of Cyber Engineering, Xidian University, 710071, Xi'an, China
[b] School of Computer Science and Information Technology, RMIT University, Australia
[c] School of Computer Science and Technology, Xidian University, 710071, Xi'an, China
[d] Noah's Ark Lab, Huawei Tech, China
[e] School of Software, Xidian University, 710071, Xi'an, China

ARTICLE INFO

Article history:
Received 7 September 2016
Revised 14 July 2017
Accepted 6 September 2017
Available online 6 September 2017

Keywords:
Hidden community
Frequent episodes
Check in
Sequence
Location

ABSTRACT

Location-based Social Networks (LBSN) such as *Foursquare* have become ubiquitous due to the vast spread of smart mobile devices. The users of LBSN can check in at any locations, write tips, share comments, *etc.* LBSN has attracted much research attention due to the fact that the large volume of check-ins provides an unprecedented channel to connect users' online world to their offline physical activities. An essential analysis task is to identify hidden community from these check-ins, *i.e.*, the community not explicitly defined based on the online links in LBSN but implicitly exist in the mobility patterns from these check-ins. The current state-of-the-art systems explore user check-ins to extract the visited venues; analyze the semantic information of the visited venues to model user behaviour patterns; cluster the users with similar patterns into the same community. They assume that users' check-ins specify the exact venues they visited. However, it has been observed that more than half of users' check-ins, *e.g.*, in *Twitter* and *Foursquare*, are venueless (*i.e.*, exact venues visited are unknown). Motivated by this observation, this work as the first attempt investigates the hidden community detection by exploring all check-ins (*i.e.*, check-ins with/without exact venues). The idea is to identify a list of nearby venues around venueless check-ins. To obtain unfailing regularity even though it is uncertain on which venues have been eventually visited, we develop methods based on the probabilistic venue sequences. By tackling a number of unique technical challenges, the rich and comprehensive information disclosed from all check-ins allows us to uncover communities with high modularity which cannot be achieved by state-of-the-art systems. Empirical study conducted on both real-world and synthetic datasets show the efficiency and effectiveness of our methods.

1. Introduction

Location Based Social Network (LBSN) has become ubiquitous due to the vast spread of smart mobile devices and specially designed applications. For example, the most popular LBSN service provider, *Foursquare*, has been reported to have

* Corresponding author.
E-mail addresses: hli@xidian.edu.cn (H. Li), ke.deng@rmit.edu.au (K. Deng), cuijt@xidian.edu.cn (J. Cui), dongzhenhua@huawei.com (Z. Dong), jfma@mail.xidian.edu.cn (J. Ma), jbhuang@xidian.edu.cn (J. Huang).

http://dx.doi.org/10.1016/j.ins.2017.09.019
0020-0255/

Table 1
An example of check-ins of *Foursquare* users.

userID	venueID	venueCat	Latitude	Longitude	Timestamp	Possible Venues
99793	-	-	37.51	127.03	2011-11-12 12:33	1023942, 1024013, 1123003, 1402032
45337	-	-	53.806	−1.5514	2011-11-13 10:25	1783991, 1174322, 1185011
99793	-	-	36.24	127.55	2011-11-12 12:33	1024478, 1124075
97992	1193023	78	32.839	−96.829	2011-11-12 18:02	-
97793	-	-	36.88	127.42	2011-11-12 18:33	1048041, 1339018

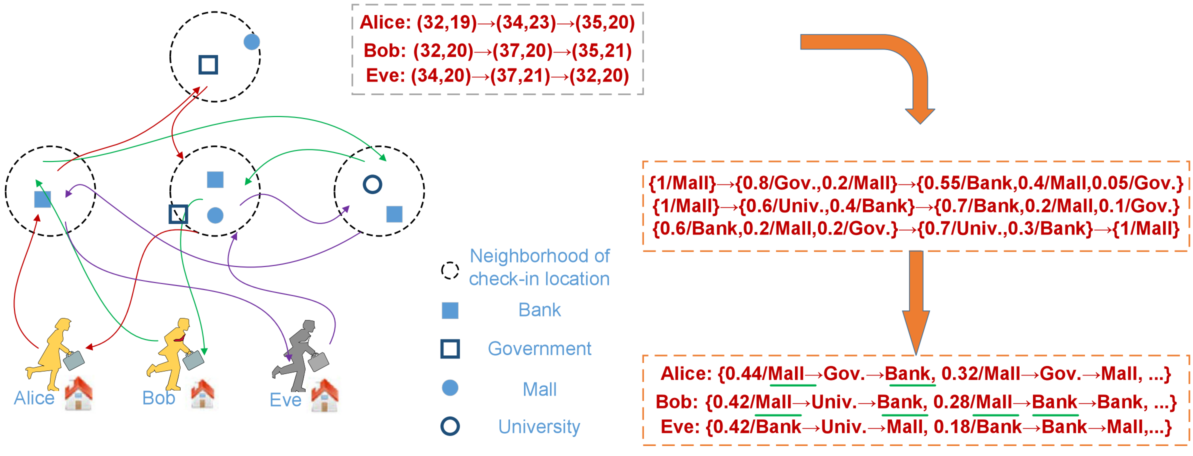

Fig. 1. Probabilistic venue sequences.

over 45 millions of registered users at the end of 2013[1]. The registered users in an LBSN service can *check in* at any locations (*e.g.,* shopping malls, hotels, restaurants, *etc.*), write tips, share comments and upload photos. The huge number of check-ins in LBSN provide us an opportunity to connect users' online world to their offline physical activities at an unprecedented level. In particular, it has been recognized that individual trajectories are not random. Instead, individual humans simply follow reproducible mobility patterns [8,15]. The information of venues frequently visited to a large extent can reveal the personal interest, consumer preference, and lifestyles of the user. It has motivated researcher to investigate the hidden community, *i.e.,* the community not directly defined based on the strength of online links as in social network [16,30,32,33] but based on the offline activities [36,37]. The hidden community detection helps identify the underlying associations between the social network users and the point of interests (POIs) in the real world which in turn enable better friends recommendation, location recommendation, and mobile viral marketing, *etc.*

Unfortunately, state-of-the-art community identification methods, even those in LBSN, suffer from the following two limitations.

On one hand, given a spatial position (*i.e.,* latitude and longitude) a check-in happens at, it may not be accurately mapped to a specific place (*e.g.,* mall, restaurant). Instead, it may correspond to a distribution of a set of places within the neighborhood of the check-in position. Table 1 shows a number of user check-ins in *Foursquare*. *venueID* is the unique identity of the recorded venue of the check-in; *venueCat* is the semantic category of the registered venue such as *restaurant, bus station, stadium, etc.; Timestamp* refers to the time when the check-in happens; *latitude* and *longitude* specify the geographical location of the check-in. In fact, some other services, including *Twitter*, also provide similar check-in functions. However, most of the location-aware check-ins in services other than *Foursquare* may not be explicitly attached with any venueID, or venueCat information as the predefined venues in these services are too limited. In fact, even in *Foursquare* itself there exist many check-ins, the types of which are "venueless"[2]; that is, the exact venue visited by the user is unknown. Alternatively, it is possible to know the nearby venues around the check-in location. This can be done by calling "search venue" function through Foursquare API by submitting the *latitude* and *longitude* of queried location. Since there is no evidence showing the exact venues the user visited, these nearby venues can be listed in field *Possible Venues* in Table 1[3]. Even though a large number of venueless check-ins exist, the current hidden community detection methods ignore them and assume all check-ins contain exact venues visited [36,37].

On the other hand, ignoring the check-in *sequential* patterns but simply comparing the frequent check-in venues between users may fail to distinguish different communities in many cases. For instance, the left side of Fig. 1 illustrates the daily

[1] http://techcrunch.com/2013/12/19/foursquare-series0d/.
[2] https://developer.foursquare.com/docs/responses/checkin.
[3] In *Twitter* case, we can also solve the venueless problem by submitting such a query through Foursquare API to form a list of *Possible Venues*.