A Diffusion Mechanism over Interest-based Communities in Mobile Internet

Yufei Di, Yuanyuan Qiao, Jie Yang, Jun Liu
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Email: yyqiao@bupt.edu.cn

***Abstract*—Mobile Internet has played an essential role in our daily life. An overwhelming majority of marketers have noticed the great opportunity to diffuse information by this platform. However, most of the diffusion mechanisms are for a certain mobile social media, which can not reflect the characteristics of innovation diffusion in whole mobile Internet. In this paper, we use a three-step method to identify the hidden relations among different users, and establish a number of interest-based communities based on the characteristics of mobile Internet. Considering the structure of these communities and user preferences, we propose a novel diffusion mechanism. Our experimental results show that the proposed model can provide an ideal effectiveness on innovation diffusion.**

Keywords*—*Mobile Internet; Interested-based communities; Innovation diffusion; Target advertisement.

I. Introduction

With the rapid development of mobile devices and wireless technologies, the traffic volume of mobile Internet has been growing continuously. The market of content delivery in mobile network is forecasted to grow to $ 5.5 billion in 2015 [1]. In 2014 December, China's mobile Internet users reached 557 million, and had an increase of 56.72 million [1]. Mobile Internet is not a substitute for traditional Internet, but an revolution of it. Such complex social ecosystems are characterized by two fundamental components, the creation of social connections between individuals and the information shared by them. Mining the static and evolutionary patterns of such phenomena is the key point to understand and predict micro and macroscopic dynamics of the whole network. Therefore, many efforts focused on investigating the causes that determine the creation of social links and the process of information diffusion along these links. On the one hand, a number of results obtained by previous work are supported by well-known sociological theories. On the other hand, most of these research works focus on some certain websites, such as Facebook, Twitter, aNobii and so on [2], [3]. However our research concentrates on finding a proper diffusion mechanism applying for the whole mobile Internet.

Those research works which focus on online social media, have great influence on the effectiveness of advertisement distribution [4]. However, the scope is limited because the connections between users are based upon exiting social relationships relying on a certain social media. As we known, the reason why mobile network has raised more and more attentions is not just about the mobile social media, but also the development of mobile video, game, and online shopping. All of them can provide an excellent platform for mobile users to share and receive information and marketers to diffuse information.

Among other researches, the diffusion of a piece of information across the mobile network still have many unexplored sides. Firstly, even though many studies have addressed the problem of predicting the global evolution of social graphs, only few investigations have been focused on the whole mobile network perspective, namely it is still not clear to what extent information can spread quickly and effectively in the network, and how to apply these models to the global mobile Internet. Similarly, even if several models of information spreading based on certain social networks have been proposed in the past, it is not applicable for mobile Internet for its limited data resource. As mentioned above, with the rapidly development, social media is not only the platform we get information from mobile Internet any more. Therefore, traditional diffusion mechanism for social advertisement over social network is not suitable for this condition.

We contribute to shed light on these questions through the analysis of a core network of a leading mobile operator in China. Unlike the mainstream, analysis with general-purpose social networks (e.g., Facebook, Twitter), which have natural relations between users, we detect communities from the data of mobile Internet based on interest, which means social aggregation is determined by the common interest websites of the mobile users. The specificity of the domain considered and the richness of the features get from the real mobile networks allow the exploration of the social dimensions from a novel perspective. Our research achives three main goals:

(1) An effective strategy for the detection of links among mobile users is designed. As opposed to the task of link detection in a general-purpose social networks, link and community detection in whole mobile networks is a task that is not well studied. Based on the survey of a large amount of strategies for community detection, we design a three-step method. With this strategy, we divide the mobile Internet users into different interest-based communities.

(2) A quantitative measure of influence in an interest-based domain by calculating the traffic usage of a user and his "friends" (someone who have a "link" to this user generated by our links detection method) is provided.

(3) A diffusion mechanism based on the interest-based community and the measure of influence is proposed. Our mechanism provide marketers a great opportunity to diffuse information through numerous populations in the whole mobile Internet.

QSHINE 2015, August 19-20, Taipei, Taiwan

DOI 10.4108/eai.19-8-2015.2261134

The rest of the paper is organized as follows: Section 2 reviews related works. Section 3 presents the research methodology and the framework of our system. Section 4 details our experiment using the three-step method, and then make some analyses and suggestions. Finally, section 6 concludes this study and describe the future work.

II. RELATED WORK

Previous works on social networks always focus on discovering patterns that describe the social feature and how social ties features evolve in time. They have revealed that link creation is driven by proximity, triangle closure, reciprocation and homophily [5]–[11]. Recently, findings from social network analysis have been corroborated and expanded by the study of communication networks [2]. Although the importance of social and communication links has been assessed in the past, and many social phenomena such as influence have been studied using the graph of conventional social ties, the effectiveness of social phenomena modeling and predicting from the view of mobile networks has not been explored thoroughly.

Besides, the tasks of capturing the dynamics of information spreading in networked environments have received much attention recently. In [12]–[14], the author has already proposed and evaluated a threshold model for the spreading of fads. Besides the simulation experiment, [2] focused on the influence phenomenon in a social context. Authors in [3] provided a diffusion mechanism for social advertising over microblogs. In this paper, we consider both networ structure and user habits to evaluate the diffusion capabilities on nodes in whole mobile Internet. We also design a diffusion mechanism which gives marketers a global guide to diffuse information through numerous populations.

Specifically, the goal of our research is to find some communities in which users are correlated by shared interest. It is much different from common social media such as Facebook, Twitter and so on, in which relations already exit in some lists of friends. We call such a set of user as an interest-based community. We propose a three-step method to identify the interest based community from mobile Internet. Based on these communities, a global diffusion mechanism for advertiser is proposed, which focuses on the whole mobile Internet rather than a specific social media.

III. METHODOLOGY AND THE SYSTEM ARCHITECTURE

In this section, we firstly describe our three-steps method for identifying user communities on the basis of users browse interest and also present an algorithm to detect most influential users. Then, a social diffusion mechanism to disseminate advertising information via influential user is designed based on these communities. In this paper, we detected communities with three steps, namely, affinity measuring, graph sparsification and communities identifying [15].

A. Affinity Measuring

The first step aims to reveal the relationships among different web users by generating a directed graph. For the purpose of building such graph, a proper factor that can reflect the common interest of two users is needed. Most services on the mobile Internet need a service provider, and different service providers use different hosts to provide their service. Usually, users connecting to different hosts indicates that different applications are used. If two users always connect to the same host, they are more likely to use the same application on smartphone, and have similar interest. Therefore, if a set of users $O = o_1, o_2, , o_n$ share a set of hosts $H = h_1, h_2, , h_n$, we define the affinity measurement function in user similarity field as:

$$aff(o_i, o_j) = \begin{cases} \frac{\mu(o_i) \cap \mu(o_j)}{\mu(o_i)} & i \neq j \\ 0 & i = j \end{cases} \quad (1)$$

where $\mu(o_i) \cap \mu(o_j)$ are the set of hosts that o_i and o_j access commonly. The denominator $\mu(o_i)$ is used to normalize the affinity measurement. From $\mu(o_i) \cap \mu(o_j) \leq \mu(o_i)$, we have $0 \leq aff(o_i, o_j) \leq 1$. Note that if $aff(o_i, o_j) = 0$, it means that there is no host that o_i and o_j both access. In contrast, $aff(o_i, o_j) = 1$ means that all the hosts that o_i accessed have been accessed by o_j too.

With the affinity measurement function, we can conduct the directed and weighted graph called *affinity graph* using the following rule:

(**RULE I**) A directed edge e_{ij} in E from o_i to o_j with weight $w_{ij} = aff(o_i, o_j)$ is established if $aff(o_i, o_j) > 0$; otherwise, no edge is constructed

Using the above rule, we get a directed and weighted graph $G = (O, E, W)$, which reflects the affinity relationships among mobile Internet users by accessed hosts.

B. Graph Sparsification

In the previous section, we have constructed an affinity graph $G = (O, E, W)$ where O represents the users, E represents the affinity relationships between nodes in O, and W is the set of weight of each edge in E. However, from the graph point of view, the full affinity graph may poll too large relations between nodes. In fact, human's cognitive ability is restricted that a man can't be friends with everybody [16]. For this property, the sparsification of the dense graph is proposed to reduce the computational workload and improve the quality of analysis [17].

Conceptually, sparsification methods can be categorized into two types, global threshold approach and nearest neighbor approach [17]. Compared with the nearest neighbor approach, the global threshold approach is more widely used in complicated situations with a large number of nodes, and it is more suitable for our dataset. The global threshold based sparsification function spa in our experiment can be formally defined as:

$$spa(o_i, o_j) = \begin{cases} 1 & aff(o_i, o_j) \geq \tau \\ 0 & aff(o_i, o_j) < \tau \end{cases} \quad (2)$$

where τ is a globe threshold. Note that we choose τ by the scale-free topological criterion method proposed in [15], [18]. The rule of scale-free topology criterion is defined as follows:

(**RULE II**) Scale-free topology criterion rule: Choose the first parameter value as the expected threshold when the saturation of fit is above 0.8.

By equation (2), we get a directed and unweighted graph

$G' = (O, E')$, where E' is the set of edges, and spa is greater than 0.

C. Communities Identifying

As mentioned before, the most important task of the proposed method is generating top-k interest based communities, that is, the k most influential users and their followers. In a graph, the influence of a node is determined based on its position in graph. Some straightforward measures of influence like degree centrality, betweenness centrality and closeness centrality have been proposed [6] [19] [5]. Although they are suitable for detecting communities by normalizing the centralities, the above indexes are insufficient to estimate the influence of one node on the whole network. Therefore, designing an effective method to identify influential nodes is necessary. Our algorithm follows two important assumptions:

(1) The influence score of a node depends on the number of nodes with which the node has close affinity relationship.
(2) The influence score of a node depends on the traffic usage of nodes with which the node has close affinity relationship.

Therefore, given sparsified affinity graph $G' = (O, E')$, the *affinity rate* from o_j to o_i is defined as:

$$affr(o_j, o_i) = \begin{cases} \frac{\mu(o_j)}{sum_{k=1}^{N} e'_{jk}} & e'_{ji} = 1 \\ 0 & e'_{ji} = 0 \end{cases} \quad (3)$$

where e' is the edges of the sparsified affinity graph G', and $\mu(o_j)$ is the traffic generated by user o_j. A *affinity rate* can be regarded as the weighted dedication of node o_j. It represents the influence of o_i added by the dedication and weight of o_j. Subsequently, we can define the *influence score* as:

$$ins(o_i) = sum_{j=1}^{n} affr(o_j, o_i) \quad (4)$$

We apply Equation [4] on G' and calculate the influence score of each node. Then, k nodes called endorsers with the k biggest influence scores are selected into S. At last, every node in S with the nodes having an edge to it are grouped together as a community.

D. System Architecture

We design a social diffusion mechanism to disseminate advertising information via influential users with our interested-based communities. The process of our diffusion mechanism is detailed as follows:

(1) We use the proposed three-step method to select top-k communities and the influential users (endorsers) of selected communities.
(2) We manually divide different hosts into different types. For example, music.qq.com belongs to music type. Certain type stamps are given to the endorsers according to the host they connect to. However, we only concentrate on popular hosts. And we define hosts that are connected by over 10 percent of users in a community as popular.
(3) The system delivers relevant advertisements to identified social endorsers.

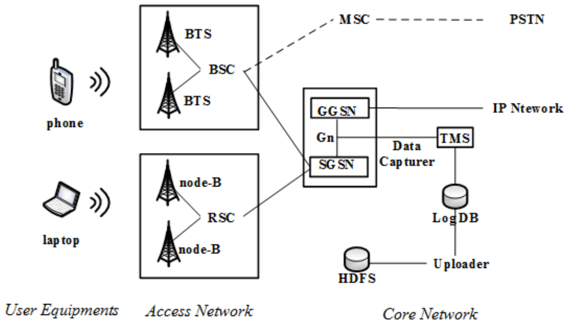

Fig. 1: Network Architecture.

IV. EXPERIMENT

A. Data Set

In this paper, the data set we use is collected by high performance network traffic monitors placed in the core network of a leading mobile operator in China. The high level view of the network structure is shown in Fig. 1. Both 2G and 3G mobile network data are collected. Our network mainly consists of three parts: (1) a number of User Equipments (UE) such as phone, pad, laptop and so on. (2) the Access Network composed of BTS/BSC (2G) and node-B/RNC (3G), and (3) the Core Network with SGSN (Serving GPRS Support Node) and GGSN (Gateway GPRS Support Node). A mobile device communicates with a base station in the Access Network which transmits its data service traffic to a SGSN. The SGSN establishes a tunnel on the Gn interface with a GGSN that provides connectivity to external networks. Through this path, a UE can connect with the IP network and reach the serving server.

Our data covers 3 million users over one week with nearly 3 billion HTTP traffic records from March 23rd to March 29th. These HTTP traffic records in our data set are saved in a log database and periodically uploaded by an Uploader component to HDFS (Hadoop Distributed File System). For the individual privacy reason, users' phone number is replaced by a sequence number, which can be used for marking subscribers and does not affect the usefulness of our analysis. Each HTTP record comprises a series of HTTP request-response events with connected host, total traffic bytes and so on.

B. Result Analysis

1) Community Detection: We make a CDF (Cumulative Distribution Function) figure of bytes usage of each user in the week in Fig. 2. We find that 80% of users only contribute 0.21% of total HTTP traffic and 70% of users generate less than 10MB HTTP traffic in the week. According to the above analysis, we apply the proposed three-step method to the selected heavy users (3000 users totally). Then, an affinity graph is constructed with Equation 1. Every node represents a user, and has a number of directed and weighted edges to other nodes.

In order to obtain the value of threshold τ, we plot a figure of the fitting index versus varied threshold τ in Fig. 3. From the figure, we can see that the scale-free fitting index have a jump when the threshold ranges from 0.49 to 0.50. According to the scale-free topology criterion rule (Rule II in Sec.3), 0.5

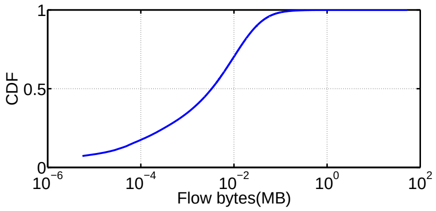

Fig. 2: CDF of Bytes Usage of Each User in One Week.

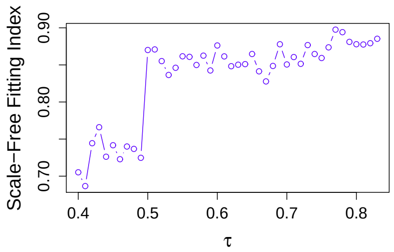

Fig. 3: Scale-Free Fitting Index

is the ideal threshold.

With the threshold $\tau = 0.5$ and Equation 2the sparse graph can be obtained. At last, we calculate the *influence score* of each node and identify the interest-based communities. Fig. 4 is an example of them.

2) Diffusion Strategy Analysis: In this part, a detailed insight into the interest-based communities will be given, to prove these communities is applicable to the diffusion mechanism we proposed.

a) Community Structure: Fig. 5 is cumulative probability distribution of links usage of each user in one week. Obviously, the interest-based networks we proposed yields a power law distribution of links one user holds. We call this kind of networks as scale-free networks.

b) Diffusion Principle: In a scale-free network, [20] have already proven the following rules:

(**RULE III**) In scale-free networks, a higher level of cost constraints for the number of neighbors of the users leads to a lower diffusion of the innovation.

(**RULE IV**) In scale-free networks of consumers, a higher level of social influence exerted by the more connected users leads to a higher diffusion of the innovation.

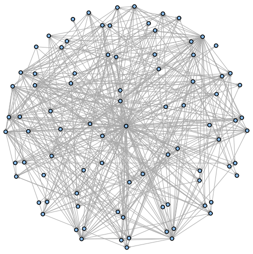

Fig. 4: Community Example

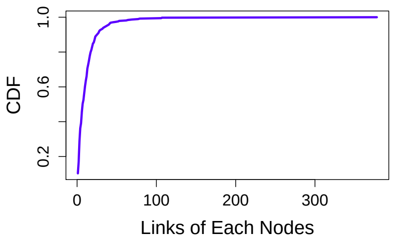

Fig. 5: CDF of links usage of each user in one week.

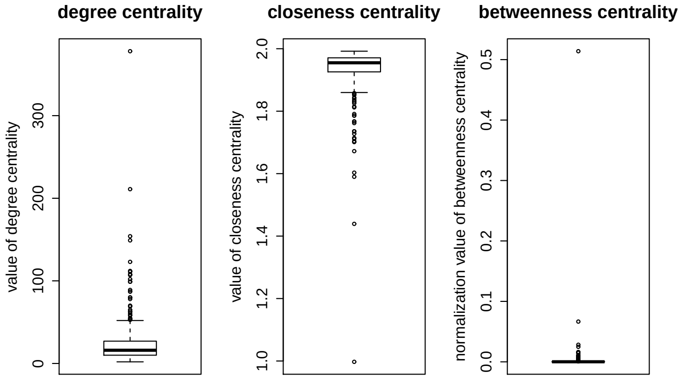

Fig. 6: Centrality of Community 1

(**RULE V**) In scale-free networks, the more the users direct their links to the endorsers, the more the innovation diffuses.

c) Effectiveness of Community Structure: We calculate the degree, closeness, and betweenness centralities of our communities. We take community 1 which have 378 users as an example to show the character of our community structure. The result is shown in Fig. 6. We can see that, our interest-based community is a central network. According to Rule III, a central network leads a higher diffusion of innovation. We deliver relevant advertisements of the sponsors to identified social endorsers because of Rule IV. And from Fig. 6, the central node of our community has links to all the users in its community, it is coincidence with Rule V.

d) Effectiveness of Endorsers Selection: Compared with the mechanism proposed in [3], which delivers advertisements to endorsers and makes endorsers' friends to be new social endorsers, our system only delivers once. Our method outperform the previous one in two aspects: (1) the structure of our community is different from traditional social media networks. In Fig. 7 we can see the in-degree and out-degree of a node are much different. One node with a high in-degree means it can influence many other nodes. On the other hand, if the out-degree is high, it indicates this node will be

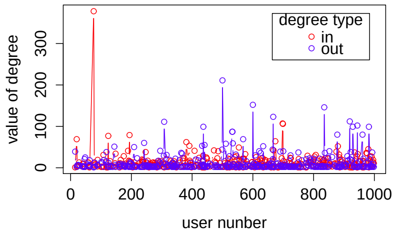

Fig. 7: Comparison of In-degree and Out-degree

influenced easily. Namely, if you have a high in-degree value, it means many other nodes have common interest to you and you can influence their browse interest according to Equation (1). Therefore, only the users who have larger in-degree can be defined as endorsers. The mechanism in [3] is not applicable to our communities. (2) Our communities are overlapping communities. That means our algorithm allows a user to join more than one groups from different perspectives. So, one user may have different endorsers in different communities. It avoids a single user has too little endorsers to accept the innovation. Above facts prove that our diffusion mechanism can work efficiently.

V. Conclusion

With the development of mobile Internet, it's important to find the hidden relations among mobile users. For service providers, having a knowledge of the user communities can help them make their marketing strategy more efficient. Especially the innovation diffusion benefits to diffuse information through numerous populations for marketers. For individuals, it is an effective approach to recommend contents they may be interested in. In this paper, we use a three-step method to identify the influential interest-based communities with the data from a large real-world cellular data network. Compared with the traditional social media network, our method can identify the hidden interest relations among different users. Utilizing these interest-based communities and the information of user behavior in mobile Internet, we design an one-time endorse diffusion mechanism, which considers user preference over mobile Internet. Experimental results show that the proposed diffusion mechanism can be effective over mobile Internet.

However, we still have a lot of work to do due to the features of our dataset. The two future works are: Firstly, we will do an effective experiment to trace the innovation diffusion among mobile users with a long term dataset. Secondly, host is not an accurate index to identify the interest of mobile users. In the future, a more accurate user profile will be given by a deep analysis of HTTP messages of mobile users.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China (61072061), Innovation Foundation of Beijing University of Posts and Telecommunications (2015RC11) , 111 Project of China (B08004) and EU FP7 IRSES MobileCloud Project (612212).

References

[1] "The 35th statistical report on internet development in china," *copyright by China Internet Network Information Center (CNNIC)*, pp. 28,75,110, 2015.

[2] L. M. Aiello, A. Barrat, C. Cattuto, R. Schifanella, and G. Ruffo, "Link creation and information spreading over social and communication ties in an interest-based online social network," *EPJ Data Science*, vol. 1, no. 1, pp. 1–31, 2012.

[3] Y.-M. Li and Y.-L. Shiu, "A diffusion mechanism for social advertising over microblogs," *Decision Support Systems*, vol. 54, no. 1, pp. 9–22, 2012.

[4] K. Lewis, M. Gonzalez, and J. Kaufman, "Social selection and peer influence in an online social network," *Proceedings of the National Academy of Sciences*, vol. 109, no. 1, pp. 68–72, 2012.

[5] T. Han, N. Ansari, M. Wu, and H. Yu, "On accelerating content delivery in mobile networks," *Communications Surveys & Tutorials, IEEE*, vol. 15, no. 3, pp. 1314–1333, 2013.

[6] A. Chin, "Finding cohesive subgroups and relevant members in the nokia friend view mobile social network," in *Computational Science and Engineering, 2009. CSE'09. International Conference on*, vol. 4. IEEE, 2009, pp. 278–283.

[7] A. Chin and M. Chignell, "Automatic detection of cohesive subgroups within social hypertext: A heuristic approach," *New Review of Hypermedia and Multimedia*, vol. 14, no. 1, pp. 121–143, 2008.

[8] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu, "Community detection in large-scale social networks," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM, 2007, pp. 16–25.

[9] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[10] B. Ball, B. Karrer, and M. Newman, "Efficient and principled method for detecting communities in networks," *Physical Review E*, vol. 84, no. 3, p. 036103, 2011.

[11] P. Bródka, S. Saganowski, and P. Kazienko, "Group evolution discovery in social networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*. IEEE, 2011, pp. 247–253.

[12] E. Abrahamson and L. Rosenkopf, "Social network effects on the extent of innovation diffusion: A computer simulation," *Organization science*, vol. 8, no. 3, pp. 289–309, 1997.

[13] Y. Wang, "System and method for targeted ad delivery," Jun. 8 2010, uS Patent 7,734,632.

[14] J. S. Hendricks, A. E. Bonner, J. S. McCoskey, and M. L. Asmussen, "Targeted advertisement using television delivery systems," Oct. 8 2002, uS Patent 6,463,585.

[15] J. Liu and N. Ansari, "Identifying website communities in mobile internet based on affinity measurement," *Computer Communications*, vol. 41, pp. 22–30, 2014.

[16] R. I. Dunbar, "Coevolution of neocortical size, group size and language in humans," *Behavioral and brain sciences*, vol. 16, no. 04, pp. 681–694, 1993.

[17] P.-N. Tan, M. Steinbach, V. Kumar *et al.*, *Introduction to data mining*. Pearson Addison Wesley Boston, 2006, vol. 1.

[18] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, 2005.

[19] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1979.

[20] S. A. Delre, W. Jager, T. H. Bijmolt, and M. A. Janssen, "Will it spread or not? the effects of social influences and network topology on innovation diffusion," *Journal of Product Innovation Management*, vol. 27, no. 2, pp. 267–282, 2010.