 CMC, vol.57, no.3, pp.571-587, 2018

Seed Selection for Data Offloading Based on Social and Interest Graphs

Ying Li[1], Jianbo Li [1, *], Jianwei Chen[1], Minchao Lu[1] and Caoyuan Li[2, 3]

Abstract: The explosive growth of mobile data demand is becoming an increasing burden on current cellular network. To address this issue, we propose a solution of opportunistic data offloading for alleviating overloaded cellular traffic. The principle behind it is to select a few important users as seeds for data sharing. The three critical steps are detailed as follows. We first explore individual interests of users by the construction of user profiles, on which an interest graph is built by Gaussian graphical modeling. We then apply the extreme value theory to threshold the encounter duration of user pairs. So, a contact graph is generated to indicate the social relationships of users. Moreover, a contact-interest graph is developed on the basis of the social ties and individual interests of users. Corresponding on different graphs, three strategies are finally proposed for seed selection in an aim to maximize overloaded cellular data. We evaluate the performance of our algorithms by the trace data of real-word mobility. It demonstrates the effectiveness of the strategy of taking social relationships and individual interests into account.

Keywords: Mobile social network, social data offloading, extreme value model, Gaussian graphical model.

1 Introduction

With the rapid progress of mobile communication technologies, it promotes massive growth of smart devices, rising popularity of mobile applications and eager engagement with connected services [Zheng, Yang, Zhang et al. (2016)]. For example, people have an active tendency to download interesting contents via mobile devices, such as multimedia newspapers, weather forecast, pictures and movies. Mobile data traffic is therefore predicated to unceasingly grow at a fast rate in accordance with ongoing tsunami of mobile data demand in the next few years. This situation poses a great challenge for future cellular network and becomes a serious concern of mobile network operators.

To increase the cellular network capacity, the operators have proposed various straightforward solutions, such as cellular network upgrade and new network construction. Moreover, mobile data/traffic offloading has attracted a lot of attentions in

[1] College of Computer Science and Technology, Qingdao University, Qingdao, 266071, China.

[2] School of Computer, Beijing Institute of Technology, Beijing, 100081, China.

[3] Faculty of Engineering and Information Technology, University of Technology Sydney, NSW 2006, Australia.

[*] Corresponding Author: Jianbo Li. Email: lijianbo@188.com.

mobile social networks (MSNs) because of avoiding overwhelming the cellular network infrastructure, see Rebecchi et al. [Rebecchi, De Amorim, Conan et al. (2015); Zhang, Chu, Guo et al. (2015); Ma, Li and Qiu (2011)]. Users can find neighbors with the same interests and build up the short-range connection with them to receive the contents, e.g., Bluetooth, Wi-Fi Direct, Near-Field-Communication (NFC), and Device-to-Device (D2D) communications. It is a low-cost solution for operators while providing low-price complementary network with high data rate for users.

One of the promising approaches for offloading mobile data/traffic is to promote opportunistic D2D communications among mobile users whenever possible, see Aijaz et al. [Aijaz, Aghvami and Amani (2013); Li, Qian, Jin et al. (2014)]. In detail, the network operators first push data items to a set of users as initial seed users vial cellular networks, and then the seed users share content with other neighboring users via short-range communications. Some users will access the cellular infrastructure to directly download the contents if they fail to receive them after a certain delay. Nevertheless, we still meet some critical challenges as follows. The first challenge we consider is how to stimulate active and effective cooperation for D2D communications among users since different individuals pursue different interests and users might more likely share with others the same data they are interested in. The second challenge we concentrate on is how to increase the opportunistic communications for content sharing among users. Last but the most important challenge is how to design the seeding strategy for drastically reducing the data traffic from the cellular network operators while satisfying the access requirement of all users? It has been studied that selecting an appropriate initial set of seeds can be a critical building block for efficient data offloading, see Wang et al. [Wang, Chen, Han et al. (2012)]. To answer these three questions, we are motivated to propose a new algorithm of choosing influential and popular seed users for content sharing via D2D communications.

The rest of the paper is prepared as follows. Section 2 briefly reviews the related work. In Section 3, we then detail the framework of our proposed algorithm, which designs modeling social properties of users to assist seed user selection. Section 4 implements experimental evaluation and analysis to verify the performance of our proposed algorithm. In Section 5, we discuss our proposed methodology and conclude the paper.

2 Related work

In this section, we review the related work of opportunistic offloading via D2D communications and the mechanism of wisely seed-set selection in the data sharing process. The final goal of all those works is to maximally reduce overloaded cellular traffic networks.

2.1 Opportunistic offloading via D2D communication

The study of opportunistic offloading via D2D communications has increased in popularity recently. Online and offline human social activities are employed to enhance D2D communications considering the fact that the handheld devices are carried by human beings and actively involved in people's routine mobility. Chen et al. [Chen, Proulx, Gong et al. (2015)] studied the cooperative D2D communications by exploring

social ties in human social networks. They designed the mechanism of social-tie-based cooperation based on two key social phenomena, social trust and social reciprocity, which enhance efficient cooperation for D2D communications among users. Wang et al. [Wang, Sun, Song et al. (2015)] proposed a strategy of user-centered opportunistic offloading in D2D-enhaneced networks. They exploit the theory of a network formation game to form a cooperative network, in which the behaviors of user selfishness are considered into account in the D2D sharing process. To maximize content sharing via D2D communication, Jiang et al. [Jiang, Zhang, Li et al. (2016)] focused on the study of selectively caching popular content and maximally matching between senders and receivers. Correspondingly, they proposed an interference-aware communication model, which formulates selective caching as a Knapsack problem and sender-receiver matching as a maximum weighted matching problem in a bipartite problem.

2.2 Seed user selection for data dissemination

The selection of high potential user as initial seeds for data dissemination significantly determines the performance of offloading mobile data traffic via opportunistic communications. Efforts have been carried out to offload traffic by the mechanism of seed user selection. Some works focus on maximizing the overall system performance to deal with data offloading problems, see Han et al. [Han, Hui, Kumar et al. (2012); Li, Qian, Jin et al. (2014)]. Han et al. [Han, Hui, Kumar et al. (2012)] study how k seed users is only selected to minimize the mobile data traffic over cellular networks. Li et al. [Li, Qian, Jin et al. (2014)] formulated the optimal seed user selection as a problem of submodular function maximization under multiple linear constraints, including traffic heterogeneity, user interests and limited storages Both of these strategies employ greedy selection algorithms to identify a sub-optimal seed set since the subset selection problem is NP-hard.

However, above-mentioned strategies commonly concentrate on user mobility without considering the practical social relationships among users. Wang et al. [Wang, Chen, Han et al. (2014)] proposed an approach of Traffic Offloading assisted by Social network services via opportunistic Sharing in mobile social networks (TOSS). The algorithm of TOSS first selected a group of seed users and pushed the same contents to them. Seed users then share contents with others via local connectivity while meeting opportunistically them. In especial, the choice of the appropriate seed users critically depends on the construction of online Social network services (SNS) and offline mobile social networks (MSNs). VIP delegation is designed to alleviate overloaded cellular networks through opportunistic offloading [Barbera, Viana, De Amorim et al. (2014)]. The basic idea is that a few but important users are selected as a bridge to transfer massive data between the network infrastructure and the remaining of the network. To guarantee that VIP users can regularly contact all the rest of the network users, the proposed method construct a social graph to leverage the VIPs.

3 Design of the new social offloading model

This section presents the mechanism of our data offloading. Section 3.1 briefly introduces the basic framework of our work. In Section 3.2, we illustrate the structure learning of

interest graph to indicate user-user interest associations. Section 3.3 constructs the contact graph, capturing social ties of users in the real world. A contact-interest graph is constructed for describing social ties and individual interests of users in Section 3.4. Relying upon the constructed three types of graphs, the three strategies are proposed for seed selection in Section 3.5.

3.1 System overview

In our daily life, people usually move with a certain purpose, affected by their interests or social roles in the real world. Besides, just a few of popular users have an "influential" role rather than the unpopular ones within the network. Motivated by these facts, we propose our new social offloading approach of choosing a few popular users as seed users based on user-user interest interactions and region-based user mobility. The seed users are capable of regularly delivering data items to them via D2D communications when the rest of users send access requirements, as displayed in Fig. 1. It is expected to greatly reduce data traffic by the use of our approach.

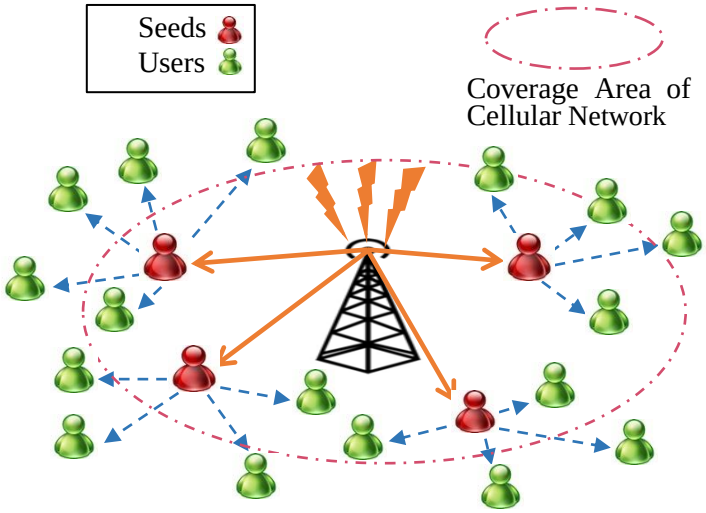


Figure 1: Illustration of our work

We first build item and user profiles to characterize heterogeneous interests of users. Gaussian graphical modeling is further employed to decode pairwise interest interactions between users and generate an interest graph. To explore social ties of users in the real word, we first threshold the pairwise contact duration into two classes by the use of extreme value theory. We think that user pairs have close relationships if theirs overall contact duration exceeds the threshold. Correspondingly, a contact graph is generated to indicate the social ties of users. Moreover, a contact-interest graph is built for charactering social ties and individual interests. We finally propose three strategies with centrality metric to select popular hubs as seed users. The wisely chosen seed users can infect a larger number of other users, resulting in drastically reducing cellular traffic.

3.2 Interest graph modeling with Gaussian graphical structure

Assume that a mobile offloading system involves in multiple data items and users. Let M={1,⋯,*n*} and N={1,⋯,*d*} be the items and users, respectively. Users usually have different interests on different items, which result in different accessing behaviors from users. We here investigate the users' interests for the different data items in this system.

It has been stated that people commonly assemble by regions and interests [Rodrigues, Benvenuto, Cha et al. (2011); Jin, Chen, Wang et al. (2013)]. Choudhury et al. [Choudhury, Sundaram, John et al. (2010)] define this phenomenon as "birds-of-a-feather" effect. Therefore, users are more willing to share and deliver the interesting items with those that have similar interests. In addition, popular data items would be more interesting to most users, while outdated data items would attract little attention. To characterize users' different interests on items, we build an item profile for each item and a user profile for each user. In this way, the complicated interest associations among users can be exposed by the use of Gaussian graphical modeling.

To create user and item profiles, a set of important features is picked to describe the preference of users and the content of data items. Let F denote the vector of features, $F = [f_1, f_2, \cdots, f_{n_F}]^T$, where n_F is the number of features. Thus, any item *i* can be well described by a vector a_i^T with length n_F, that is, $a_i^T = [a_{i1}, a_{i2}, \cdots, a_{i,n_F}]^T$. For any user *j*, the profile is built by the use of Rocchio algorithm.

Assume that

$$u_j^T = \beta \cdot \frac{1}{|M_r|} \sum_{a_s \in I_r} a_s^T - \gamma \cdot \frac{1}{|M_{nr}|} \sum_{a_r \in I_{nr}} a_t^T \tag{1}$$

where M_r denotes a set of data items that user *j* has more concerns on and M_{nr} denotes a set of data items that user *j* has less concerns on. The parameters β and γ are the positive and negative feedback weights, respectively.

Given user profile and item profile, the preferences of user j to data item *i* can be estimated,

$$x_{ij} = \frac{a_i^T u_j}{\|a_i\| \|u_j\|} \tag{2}$$

where $i = 1, \cdots, n$ and $j = 1, \cdots, d$. The matrix $X = \left(x_{ij}\right)_{n \times d}$ summaries users' interest on items.

Depending on the interest matrix *X*, Gaussian graphical modeling is then employed to explore conditional interest dependences between users. In detail, the problems are defined as follows: let G=(V, E) be an undirected graph, where *V* is the defined set of users above and $E \subset V \times V$ is the set of existing undirected edges. A Gaussian graphical model with respect to the graph G is defined,

$$\mathcal{M}_G = \{N_d(u, \Sigma) | \Theta = \Sigma^{-1} \in \mathbb{P}_G\} \tag{3}$$

where Θ is the precision matrix and $\mathbb{P}_G$ represents the space of $d \times d$ positive definite matrices with elements (*i*,*j*) equal to zero whenever $(i, j) \notin E$. The graph structure is determined by the precision matrix, $\Theta = \left\{\{i, j\} \in E; \Sigma_{ij}^{-1} \neq 0\right\}$.

Accordingly, the log-likelihood function is written as,

$$\sum_{i=1}^{n} -\frac{1}{2}(x_i - u)^T \ \Theta(x_i - u) + \frac{n}{2} logdet(\Theta) - \frac{nd}{2} log(2\pi) \quad (4)$$

where x_i is the i^{th} row of matrix X. For a given Θ, the optimal u is the mean of samples $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The ℓ_1-penalized MLE framework of the graphical Lasso is used to infer the graph structure [Friedman, Hastie and Tibshirani (2008); Witten and Friedman (2011); Zhao, Liu, Roeder et al. (2012)].

$$\hat{\Theta}^{\lambda} = argmin_{\Theta>0}\left\{\sum_{i=1}^{n} -\frac{1}{2}(x_i - \bar{x})^T \ \Theta(x_i - \bar{x}) + \frac{n}{2} logdet(\Theta) - \frac{n}{2}\lambda \sum_{i<j}\left|\Theta_{ij}\right|\right\} \quad (5)$$

Let S be the empirical covariance matrix, the equation above becomes

$$\hat{\Theta}^{\lambda} = argmin_{\Theta>0}\left\{-log\{det(\Theta)\} + tr(S\Theta) + \lambda \sum_{i<j}\left|\Theta_{ij}\right|\right\} \quad (6)$$

We here apply the method of rotation information criterion for variable selection and Gaussian graphical modeling [Wei, Duan, Shi et al. (2013)]. Stability selection integrates subsampling with high-dimensional selection algorithm. The use of stability selection method is beneficial to generate finite sample control for familywise errors and significantly improved the accuracy of structure estimation. Finally, the inferred graph is obtained as

$$\hat{\Theta}^{\lambda} = \left\{(i,j) \in E; \hat{\Theta}_{ij}^{\lambda} \neq 0\right\} \quad (7)$$

By the use of the Gaussian graphical modeling to discover interest associations between users, the inferred interest graph is elected with displaying user cliques and hub users. Users are grouped into different cliques based on similar interests. In each clique, hub users are the ones who own more links with others. Therefore, these hub users are supported as the candidates for seed users.

3.3 Contact graph construction for social relationship determination

It has been investigated that mobile users have different mobility patterns since people in real life have different levels of popularity, interests, social relationships and so on [Ma Li and Qiu (2011); Zhuo, Gao, Cao et al. (2014); Wang, Chen, Han et al. (2014)]. However, the mobility traces of users are not random but regularly follow fixed patterns. People often move toward some certain destinations, which are determined by their social roles and interests. For example, friends meet each other more frequently or stay longer than strangers. Therefore, users with close relationships have more opportunities to share items between them. In addition, users are selfish, only considering their individual payoff rather than the performance of the overall cellular traffic system. For example, selfish users can't bear sharing contents with strangers considering their own limited device storage, energy usage and private safety. Comparatively, close pair of users deserves social trust with each other and voluntarily promote mutual cooperation.

We are motivated to convert user mobility traces into a weighted contact graph for exploring social relationships. The nodes of the graphs represent users from the traces and the edges are the contacts between users. The weight of the edges is the value based on a metrics to reveal the social relationships. Scarrott et al. [Scarrott and MacDonald

(2012)] stated that contact duration indicates familiarity. In particular, a threshold approach was used to detect familiar pairs. An edge is added if the total contact duration exceeds a threshold. We here measure the relationship between users by how long they stay with each other. If two users spend more time together, they are in closer relationships. Assume that $\tilde{G} = (V, \tilde{E})$ be the contact graph, where V is the defined set of users above, $\tilde{E} \subset V \times V$ is the set of existing undirected contact edges.

We start with analyzing the distribution of total contact durations of all pairwise users in order to detect an appropriate threshold for separating close-relationship pairs from other pairs. To this end, extreme value modeling is applied for describing the likelihood of behavior of a small amount of close-relationship pairs and then choosing the threshold for distinguishing social ties. Various kinds of approaches have been developed for the choice of threshold, such as parametric, semi-parametric and non-parametric models. Considering the complicated multi-modality of the distribution of total contact duration, we explore non-parametric kernel mixture models to estimate the structure of the duration data without assuming a particular parametric form. This model equips a standard kernel density estimator as the bulk with GPD tail model [Hu (2013); Qin, Zhu, Zhu et al. (2016)]. Assume that the nonzero duration time of user pairs is composed of n_t independent and identically distributed observations $Y = \{y_1, \cdots, y_{N_p}\}$. The corresponding distribution function of kernel GPD model is defined,

$$F(y|Y, h, \delta, \sigma_\delta, \xi, \Phi_\delta) = \begin{cases} (1 - \Phi_\delta)\frac{H(x|X,\lambda)}{H(u|X,\lambda)} & y \le \delta \\ (1 - \Phi_\delta) + \Phi_\delta G(x|\delta, \sigma_\delta, \xi) & y > \delta \end{cases} \tag{8}$$

where $H(\cdot\,|X, \lambda)$ is the distribution function of the kernel density estimation, $G(x|\delta, \sigma_\delta, \xi)$ is unconditional GPD function and Φ_δ is the probability of being above the threshold δ, $\Phi_\delta = P(Y > \delta)$.

Specifically, $H(y|X, h)$ is denoted as,

$$H(y|Y, \lambda) = \int_{-\infty}^{y} p_{n_t,\lambda}^K(u)du \tag{9}$$

$$p_{n_t,h}^K = \frac{1}{n_t\lambda}\sum_{i=1}^{n} K(\frac{y - y_i}{h}) \tag{10}$$

where $p_{n_t,h}^K$ is the kernel estimator, λ is the bandwidth, $K(y)$ is a kernel function that usually meets the following conditions: $K(y) \ge 0$ and $\int K(y)dy = 1$.

Given the threshold δ, the excess $y - \delta$ can be well approximated by a generalized Pareto distribution, which is

$$G(y|\delta, \sigma_\delta, \xi) = Pr(Y < y|Y > \delta) = \begin{cases} 1 - \left[1 + \xi\left(\frac{y-\delta}{\sigma_\delta}\right)\right]_+^{-1/\xi} & \xi \ne 0 \\ 1 - exp\left[-\left(\frac{y-\delta}{\sigma_\delta}\right)\right]_+ & \xi = 0 \end{cases} \tag{11}$$

where $y > u$, $z_+ = \max(z, 0)$, ξ and $\sigma_\delta > 0$ are the shape and scale parameters, respectively. We then fit the kernel GPD models to obtain the estimated threshold δ.

In real world, user mobility guided by their interests or social roles generate repeatability in their behaviors, such as go to work/school every day, go shopping with friends. Intuitively, observing patterns of contact and interest indicates enough useful information

for the choice of seed users since users with closely social ties or similar interest might be willing to share and transfer the interesting information and resources with each other. Image a scenario that if a user is strange with others and also is not interested in the data, the processing of receiving the data via direct cellular links and delivering it to others becomes a burden for him without any gain. In this situation, selecting this user as a seed would be challenging to offload the network traffic.

We are encouraged to construct a contact-interest graph, indicating closely social relationships and similar interests of users. Details are described as follows. Users can be divided into three types based on social relationships and interests. The first one is that user pairs meet each other in a longer duration, therefore, they are viewed as being in a close relationship, such as friends, family. These users are willing to share items with each other without considering the interest factor. We naturally think that if pair of users is in this social relationship, items can be delivered with a higher probability p_1. The second one is that pair of users meet each other in a certain frequency and also have common interests. Assume users with this type of relationship share data items with a probability p_2. The third one is that pair of users encounter each other but have different interests. Assume users with this type of relationship share data items with a probability p_3. It can be obviously concluded that $p_3 < p_2 < p_1$. Under this environment, contact-interest graph can be constructed by the combination of contact graph and interest graph.

3.4 Contact-interest graph construction by considering social ties and individual interests

In real world, user mobility guided by their interests or social roles generate repeatability in their behaviors, such as go to work/school every day, go shopping with the same friends. Intuitively, observing patterns of contact and interest indicates enough useful information for the choice of seed users since users with closely social ties or similar interest might be willing to share and transfer the interesting information and resources with each other. Image a scenario that if a user is strange with others and also is not interested in the data, the processing of receiving the data via direct cellular links and delivering it to others becomes a burden for him without any gain. In this situation, selecting this user as a seed would be challenging to offload the network traffic.

We are encouraged to construct a contact-interest graph, indicating closely social relationships and similar interests of users. Details are described as follows. Users can be divided into three types based on social relationships and interests. The first one is that user pairs meet each other in a longer duration, therefore, they are viewed as being in a close relationship, such as friends, family. These users are willing to share items with each other without considering the interest factor. We naturally think that if pair of users is in this social relationship, items can be delivered with a higher probability p_1. The second one is that pair of users meet each other in a certain frequency and also have common interests. Assume users with this type of relationship share data items with a probability p_2. The third one is that pair of users encounter each other but have different interests. Assume users with this type of relationship share data items with a probability p_3. It can be obviously concluded that $p_3 < p_2 < p_1$. Under this environment, contact-interest graph can be constructed by the combination of contact graph and interest graph.

3.5 Seed user selection mechanism

Given a data item of size *M* during a time period, the total size of the offloaded data that a user can deliver to its neighbors is defined as

$$O(u_i) = \sum_{u_k \in \mathcal{N}(u_i)} M p_{i,k} S_{i,k} \tag{12}$$

where $\mathcal{N}(u_i)$ is the direct neighbors of user u_i, p_{ik} is the sharing probability of data items between user u_i and user u_k by D2D communication, and $S_{i,k}$ is the indicator of opportunistic contact of user u_i and user u_k. $S_{i,k} = 1$ if users u_i and u_k stay in proximity; otherwise, $S_{i,k} = 0$. Seed users are selected by yielding the highest offloaded data,

$$\max_{u_i} O(u_i) \tag{13}$$

However, it has proved that it is a NP-hard problem of choosing a given number of user seeds to delivery data items with maximum coverage [Clauset, Newman, Moore et al. (2004)]. Therefore, three heuristic solutions of seed selection are designed for maximizing user satisfaction as well as reducing cellular traffic.

The three proposed strategies select seeds based on community structures in an aim to improve opportunistic sharing efficiency. For instance, users in other communities the data cannot be delivered to other if most of selected seeds belong to the same community. A specified example is displayed for illustration. Assume that most of selected seeds belong to the same community. In this environment, it indeed improves the efficiency of opportunistic communication of this community. However, the users in this community can repeatedly receive the same data from different seeds and users in other communities would have little opportunity to obtain it. Consequently, the total efficiency of data offloading cannot be enhanced by this approach of seed selection. Instead, a better mechanism is to select seeds from the disjointed communities.

3.5.1 Seed selection based on contact graph

Seed users are selected based on offline contact graph, which aim to maximize opportunistic item delivery. To do this, three key steps are executed as follows. We first discover the dense subgraph, also called communities, based on the weighted contact graph. An algorithm of fast greedy optimization of modularity is implemented for detecting the community structure [Landherr, Friedl, Heidemann et al. (2010)]. Pairs with social relationship often meet with each other; therefore, these pairs have a tendency to lie in the same community. We then define the importance of a user in each community by the use of several structural attributes, including degree centrality, betweenness centrality, closeness centrality, PageRank, and eigenvector centrality. All of these attributes are well-known in social network theory [Landherr, Friedl, Heidemann et al. (2010); Brin and Page (1998)]. Five groups of seed users are selected on the basis of the importance of users, which are ranked by each attribute. Finally, we choose the group of seed users that is able to maximize the offload data items.

3.5.2 Seed selection based on interest graph

People with similar interests like to share and transfer the interesting information with each other. Therefore, we here also employed the algorithm of fast greedy optimization

of modularity to detect the structure of interest-based community. Users in the same community are often interested in the same data items. Similarly, five attributes are individually used to rank the importance of users in detected community structure. The users with high importance are probably capable of propagating information to a maximum number of "interested" users. Correspondingly, it generates five groups of seed users. We choose the group of seed users that reduces the maximal offloading data items.

3.5.3 Seed selection based on contact-interest graph

This approach takes users of social relationships and interest into consideration to construct a graph of the contact expectation according to Eq. (12). We do categorize users into community, of which users possess similar interests and locally social contacts. Under this situation, the most popular hub of each smaller group as seed users is expected to push more data items via direct D2D communications, such as, WIFI, Bluetooth and so on. This is because the popular hubs own the data items attracting other users; meanwhile have more social ties-based trust with others. Users would be free to obtain interested items without hesitation.

4 Simulations

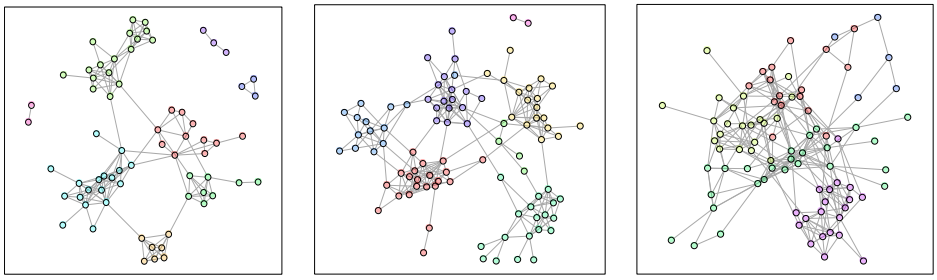
In this section, we report the performance evaluation of our proposed algorithms. We, for illustration, conduct evaluations on the trace data of real-word mobility of Reality and Infocom2006 [Scott, Gass, Crowcroft et al. (2009); Eagle and Pentland (2006)]. Reality dataset is collected by the MIT Reality Mining Project, which records contacts among 100 participants of students and staff using Bluetooth device at MIT. Infocom2006 dataset is the human mobility contact trace of Infocom conference. Participants with 79 short range and 20 long range devices were selected to generate the connectivity trace. The node with ID 99 in the original trace actually did not encounter any other users, therefore, the total number of participants was reduced to 98.

To evaluate the proposed three algorithms of seed selection, we need MSN trace data to discover the mobility patterns among users, as well as online shopping data to analyze user interests. However, we find no publicly available trace data that contains both user mobility patterns and individual interests. To deal with this issue, we simulate the user interest with the number of data items *M*=200. A set of 100 features is extracted for charactering data items, $\mathrm{F} = [F_1, F_2, \cdots, F_{100}]^T$. For data item i, the normal distribution is employed to generate its profile, $a_i^T = [a_{i1}, a_{i2}, \cdots, a_{i,100}]^T$. Considering that users often assemble by their interests in real world, we randomly arrange users into different groups. Each group of users has their own Points-Of-Interest (POIs). For the user u_j in a group, two samples are randomly selected as his interesting data items. We then calculate the user profile for u_j according to (1).

We simulate a scenario of data sharing with the number of 10 data items, $\{D_1, \cdots, D_{10}\}$ and the average size of data items with 100 M. The delay of data dissemination is considered by a constraint that the data item is delivered before its deadline. Assume data lifetimes are uniformly generated as $\{T_1, \cdots, T_{10}\}$. In addition, the parameters of p_1, p_2 and p_3, are initialized as p_1= 0.9, p_2 = 0.7 and p_3 = 0.1.

4.1 Results with reality dataset

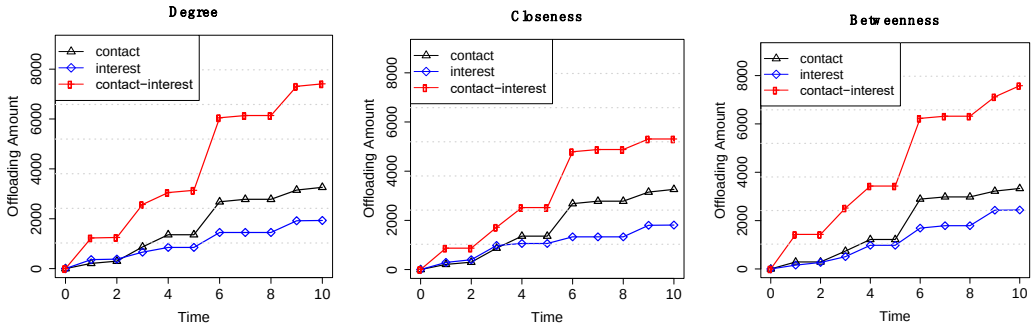
As discussed in Section 3, we exploit the movement patterns of users and social contacts among them over an observation period. The duration threshold is first calculated by the non-parametric extreme value model in order to construct the contact graph. We then conduct the interest simulation by selecting 200 popular data items and built the interest graph. The combination of contact graph and the interest graph generates the contact-interest graph. We implement community detection by the use of the fast greedy algorithm. Fig. 2 shows the detected communities for the three constructed graphs of contact graph, interest graph and contact-interest graph from (a)-(c). Different community structures are identified in the three different graphs.



(a) Contact graph (b) Interest graph (c) Contact-interest graph

Figure 2: The three graphs show the detected communities, which are denoted as different colors in the original manuscript

We investigate the offloaded data of the three strategies over time. In this scenario, the number of seeds is initialized as 15. We consider each strategy with five different centrality attributes, including Degree, Closeness, Betweenness, PageRank, Eigenvector, as shown in Fig. 3. For the contact-interest approach, the attribute of PageRank centrality has the best performance, and ultimately offloads over 8000 M data. In comparison, the offloading amount is lower than 3000 M by the use of any attribute of the contact-based and interest-based strategies. Experimental results demonstrate that seed selection for offloading data would not be efficient while only considering only single factor of social relationship and interest. Therefore, taking into account both of two factors can maximally reduce the data in the overloaded network.



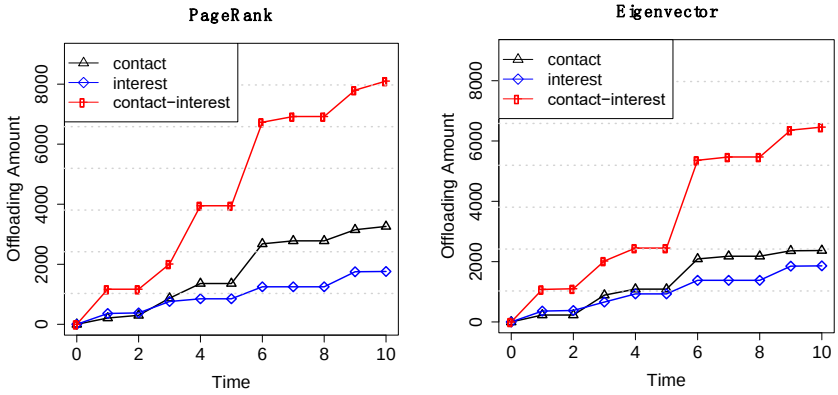
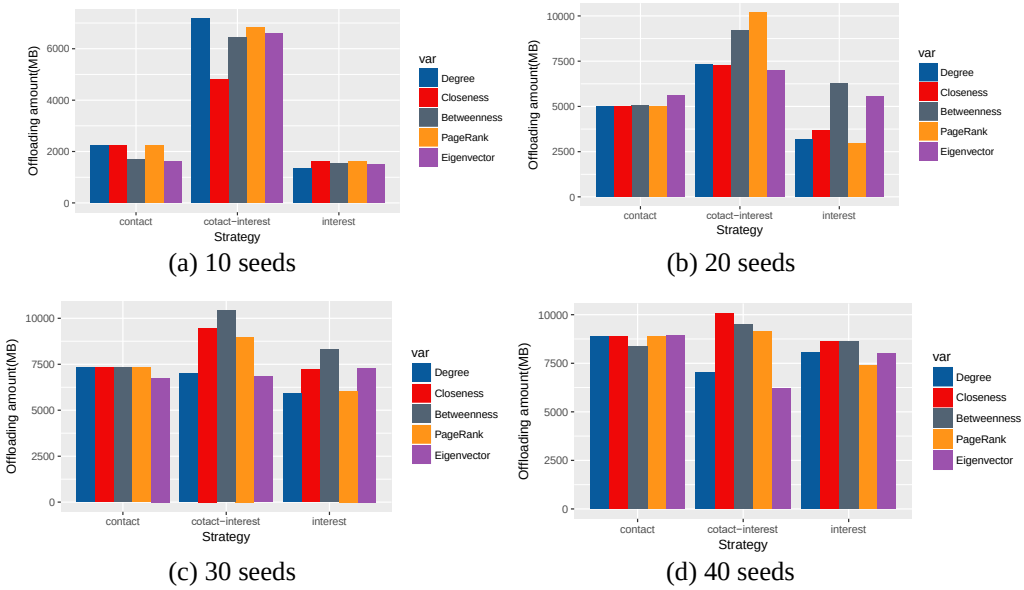


Figure 3: Performance comparison of the three proposed strategies in terms of the five different attributes over time

The number of seeds is studied in the following scenarios. On the one hand, it is not easy to realize offloading data via selecting few seeds. On the other hand, a higher number of seeds may increase the network overhead, as well as involving a high cost. We here construct four scenarios in accordance with different seed numbers in order to discover the appropriate number of seeds (Fig. 4). From (a)-(d), the number of seeds is 10, 20, 30 and 40, respectively. With the increasing of the number of seeds, it is coupled with the high amount of offloaded data for the threes strategies. It can also be easily found that the contact-interest strategy outperforms other rivals for data offloading. The attributes of closeness, betweenness and PageRank have an advantage over the attributes of degree and event for contact-interest strategy. In addition, the contact-based strategy performs comparable with interest-based strategy.

From the analysis above, the contact-interest based strategy is the best performing for realizing offloading maximization. The next critical step is to determine the number of seeds. Fig. 5 displays the amount of offloaded data varying with the number of seeds. It can be found that the contact-interest based strategy performs well with the attributes of closeness, betweenness and PageRank, which is consistent with the previous result. The use of these three attributes maximally offloads 10169 M, 10420 M, and 10290 M while the number of seeds is arranged to 26, 19, and 12, respectively. Of these three attributes, the closeness attribute is inferior because of the large number of seeds. The betweenness attribute can be applied for reducing the highest amount of cellular data, however, it requires a relatively number of seeds. Comparatively, the data can be offloaded highly with PageRank attribute while selecting few seeds. Therefore, the better choice is PageRank attribute in comparison with others.



(a) 10 seeds (b) 20 seeds

(c) 30 seeds (d) 40 seeds

Figure 4: Performance comparison of the three proposed strategies with five attributes in terms of different number of seeds for Reality dataset

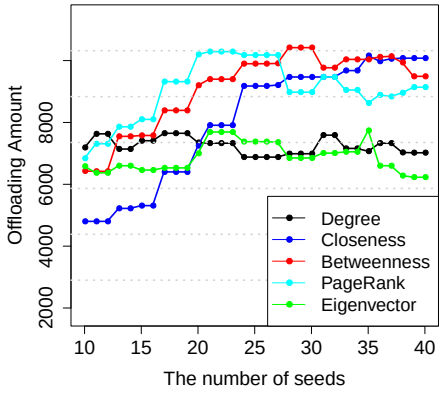
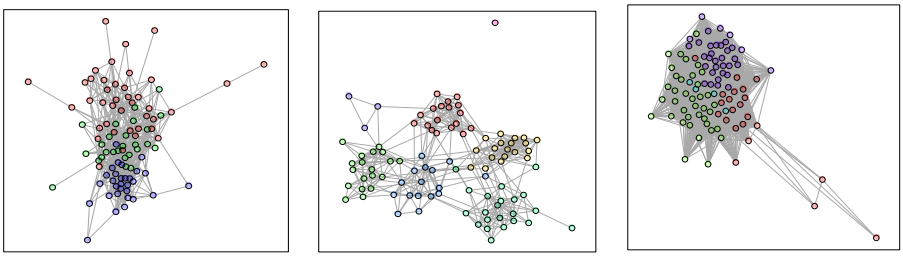


Figure 5: Determination of the number of seeds for the contact-interest strategy

4.2 Results with infocom2006 dataset

The performance of the three strategies is also investigated by exploring Infocoom2006 dataset. We first calculate the duration threshold as 9100 with the non-parametric extreme value model, which is a compromise in comparison with 7200 [Scarrott and MacDonald (2012)] and 20,000 [Zhuo, Gao, Cao et al. (2014)]. So, the close relationship can be discovered for constructing the contact graph. We here choose 200 data items related with the participants' research work for the simulation of user interest, such as books, papers, and so on, on which the interest is successfully built. The contact-interest graph is generated based on the contact graph and the interest graph. The fast greedy

algorithm is used for detecting communities from these three graphs. Fig. 6 shows the detected communities for contact graph, interest graph and contact-interest graph from (a)-(c). It can be found that the detected communities in the contact-interest graph are more close to those on contact graph. This situation is probably generated by the fact of high encounters among participants during conference. The recurrence of contacts between users can be guided by the social and interest factors, such as, the country of origin, research topic.

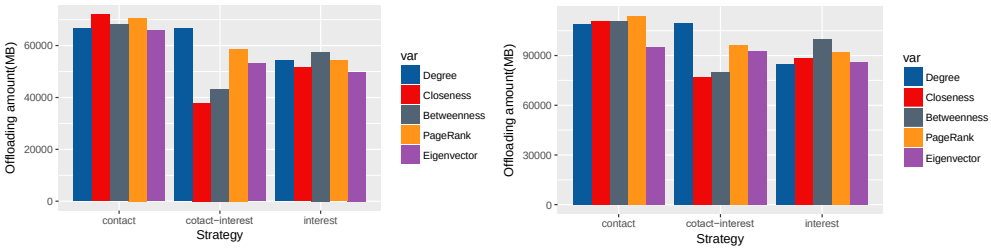


(a) Contact graph (b) Interest graph (c) Contact-interest graph

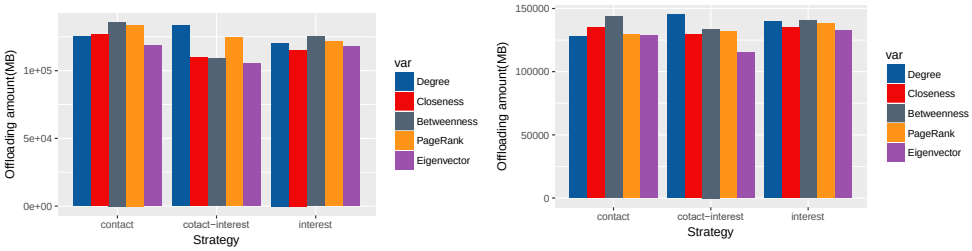
Figure 6: The three graphs show the detected communities, which are denoted as different colors in the original manuscript

The three strategies with five attributes are then considered for data offloading. We observe the performance of the tree strategies while varying the number of seeds, as displayed in Fig. 7. From (a)-(d), the number of seeds is 10, 20, 30 and 40, respectively. The contact-based strategy is a better choice while selecting few seeds. Besides, the offloading results are comparable by the use of the five attributes. The contact-interest strategy also performs well with degree attribute, although other four attributes have the inferior performance. Especially, the contact-interest strategy with degree attribute outperforms other rivals for the scenario of 40 seeds. Compared with other two strategies, the interest-based strategy is not an appropriate choice for data offloading.

We then explore the determination of the number of seeds. Considering the above-analyzed comparability of the contact and contact-interest strategies, both of the offloading results are displayed from (a)-(b) in Fig. 8. It can be easily found that it is competitive for the contact strategy with betweenness attribute and the contact-interest strategy. We naturally think that participants in the conference encounter each other more frequently. Under this situation, the high amount of cellular can be offloaded by the contact-based strategy.

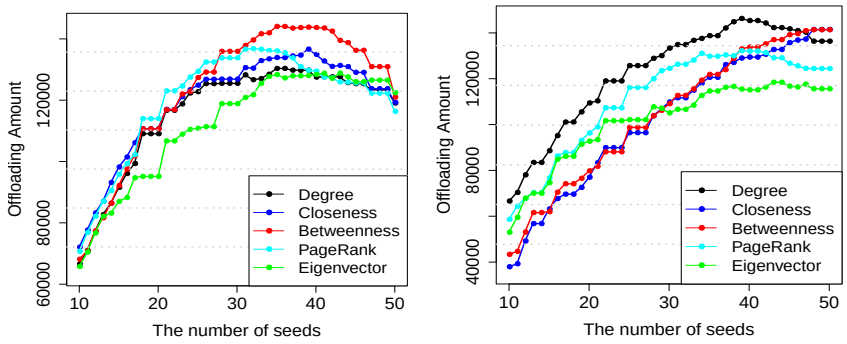


(a) 10 seeds (b) 20 seeds



(c) 30 seeds (d) 40 seeds

Figure 7: Performance comparison of the three proposed strategies with five attributes in terms of different number of seeds for infocom2006 dataset



(a) contact strategy (b) contact-interest strategy

Figure 8: Determination of the number of seeds for the contact and contact-interest strategies

5 Conclusion

In real world, user mobility generates repeatability in their behaviors, affected by their interests or social roles. We therefore have studied the social data offloading based on the exploration of social network and individual interests of users. In detail, a threshold approach is used for discovering social ties, on which a contact graph is constructed. An interest graph is also built on the basis of Gaussian graphical modeling. We then construct a contact-interest graph by the combination of interest graph and close-relationship graph. Depending on different constructed graphs, three strategies are proposed to select socially important users as seeds for opportunistically offloading overloaded data in cellular network. Experimental results indicate the usefulness of considering two factors of social ties and individual interests. Our future work will investigate the incentive mechanism with the integration of our proposed strategies in attempt to enhance offloading amount of cellular data.

Acknowledgement: This work was supported in part by National Natural Science Foundation of China under Grant No. 61502261, 61572457, 61379132, Key Research and Development Plan Project of Shandong Province under Grant No. 2016GGX101032

and Science, Technology Plan Project for Colleges and Universities of Shandong Province under Grant No. J14LN85 and the Natural Science Foundation of Shandong Province under Grant No. ZR2017PF013.

References

Aijaz, A.; Aghvami, H.; Amani, M. (2013): A survey on mobile data offloading: technical and business perspectives. *IEEE Wireless Communications*, vol. 20, no. 2, pp. 104-112.

Barbera M. V.; Viana A. C.; De Amorim M. D.; Stefa, J. (2014): Data offloading in social mobile networks through VIP delegation. *Ad Hoc Networks*, vol. 19, pp. 92-110.

Brin, S.; Page, L. (1998): The anatomy of a large-scale hyper textual web search engine. *International World Wide Web Conferences*, pp. 107-117.

Chen, X.; Proulx, B.; Gong, X.; Zhang, J. (2015): Exploiting social ties for cooperative D2D communications: A mobile social networking case. *IEEE/ACM Transactions on Networking*, vol. 23, no. 5, pp. 1471-1484.

Choudhury, M. D.; Sundaram, H.; John, A.; Seligmann, D. D.; Kelliher, A. (2010): Birds of a feather: Does user homophily impact information diffusion in social media? *Computers and Society*, pp. 1-31.

Clauset, A.; Newman, M. E.; Moore, C. (2004): Finding community structure in very large networks. *Physical Review E*, vol. 70, no. 6, pp. 1-6.

Eagle, N.; Pentland, A. (2006): Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255-268.

Friedman, J.; Hastie, T.; Tibshirani, R. (2008): Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, vol. 9, no. 3, pp. 432-441.

Han, B.; Hui, P.; Kumar, V.; Marathe, M. V.; Shao, J. et al. (2012): Mobile data offloading through opportunistic communications and social participation. *IEEE Transactions on Mobile Computing*, vol. 11, no. 5, pp. 821-834.

Hu, Y. (2013): *Extreme Value Mixture Modelling: An R Package and Simulation Study. (Ph.D. Thesis)*. University of Canterbury, New Zealand.

Jiang, J.; Zhang, S.; Li, B.; Li, B. (2016): Maximized cellular traffic offloading via device-to-device content sharing. *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 82-91.

Jin, L.; Chen, Y.; Wang, T.; Hui, P.; Vasilakos, A. V. (2013): Understanding user behavior in online social networks: A survey. *IEEE Communications Magazine*, vol. 51, no. 9, pp. 144-150.

Landherr, A.; Friedl, B.; Heidemann, J. (2010): A critical review of centrality measures in social networks. *Web Intelligence*, vol. 2, no. 6, pp. 371-385.

Li, Y.; Qian, M.; Jin, D.; Hui, P.; Wang, Z. et al. (2014): Multiple mobile data offloading through disruption tolerant networks. *IEEE Transactions on Mobile Computer*, vol. 13, no. 7, pp. 1579-1596.

Ma, L.; Li, W.; Qiu, X. (2011): Policy based traffic offload mechanism in H(e)NB subsystem, *Sia-pacific Network Operations and Management Symposium*, pp. 1-6.

Qin, J.; Zhu, H.; Zhu, Y.; Zhu, Y.; Lu, L. et al. (2016): POST: Exploiting dynamic sociality for mobile advertising in vehicular networks. *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 6, pp. 1770-1782.

Rebecchi, F.; De Amorim, M. D.; Conan, V.; Passarella, A.; Bruno, R. et al. (2015): Data offloading techniques in cellular networks: A survey. *IEEE Communications Surveys and Tutorials*, vol. 17, no. 2, pp. 580-603.

Rodrigues, T.; Benvenuto, F.; Cha, M.; Gummadi, K.; Almeida, V. (2011): On Word-of-Mouth based discovery of the Web. *Internet Measurement Conference*, pp. 381-396.

Scarrott, C. J.; MacDonald, A. (2012): A review of extreme value threshold estimation and uncertainty quantification. *Revstat-Statistical Journal*, vol. 10, no. 1, pp. 33-59.

Scott, J.; Gass, R.; Crowcroft, J.; Hui, P.; Diot, C. et al. (2009): CRAWDAD dataset: Cambridge/haggle/imote/infocom2006.
http://crawdad.cs.dartmouth.edu/cambridge/haggle/imote/ infocom2006.

Wang, T.; Sun, Y.; Song, L.; Han, Z. (2015): Social data offloading in D2D-enhanced cellular networks by network formation games. *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 7004-7015.

Wang, X.; Chen, M.; Han, Z.; Kwon, T.; Choi, Y. (2012): Content dissemination by push and share in mobile cellular networks: An analytical study. *Mobile Ad-hoc and Sensor Systems*, pp. 353-361.

Wang, X.; Chen, M.; Han, Z.; Wu, D. O.; Kwon, T. T. (2014): TOSS: Traffic offloading by social network service-based opportunistic sharing in mobile social networks. *International Conference on Computer Communications*, pp. 2346-2354.

Wei, K.; Duan, R.; Shi, G.; Xu, K. (2013): Distribution of inter-contact time: An analysis-based on social relationships. *Journal of Communications and Networks*, vol. 15, no. 5, pp. 504-513.

Witten, D.; Friedman, J. (2011): New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, vol. 20, no. 4, pp. 892-900.

Zhang, H.; Chu, X.; Guo, W.; Wang, S. (2015). Coexistence of Wi-Fi and heterogeneous small cell networks sharing unlicensed spectrum. *IEEE Communications Magazine*, vol. 53, no. 3, pp. 158-164.

Zhao, T.; Liu, H.; Roeder, K.; Lafferty, J. D.; Wasserman, L. (2012): The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1059-1062.

Zheng, K.; Yang, Z.; Zhang, K.; Chatzimisios, P.; Yang, K. et al. (2016): Big data-driven optimization for mobile networks toward 5G. *IEEE Network*, vol. 30, no. 1, pp. 44-51.

Zhuo, X.; Gao, W.; Cao, G.; Hua, S. (2014): An incentive framework for cellular traffic offloading. *IEEE Transactions on Mobile Computing*, pp. 541-555.