Data Collection through Device-to-Device Communications for Mobile Big Data Sensing

Hanshang Li Ting Li Xinghua Shi Yu Wang
College of Computing and Informatics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

***Abstract*—The appearance of smart mobile devices with communication, computation, and sensing capability and increasing popularity of various mobile applications have caused the explosion of mobile data recently. In the same time, mobile big data sensing has been emerging as a new sensing paradigm where vast numbers of mobile devices are used for sensing and collecting huge amounts of mobile data in cities. One of the challenges faced by mobile sensing is how to efficiently collect the huge amount of mobile data beyond the existing capacity of 4G networks. In this paper, we investigate the feasibility of collecting data packets from mobile devices through device-to-device communications by carefully selecting the subset of relaying devices. We formulate the problem as an optimization problem and propose a simple solution to solve it. Our experiments over a real-life mobile trace confirm the effectiveness of the proposed idea.**

I. Introduction

With the increasing popularity of mobile applications and services for smart devices, we are currently facing the challenges of mobile big data explosion. Based on the most recent Cisco's report [1], mobile data traffic grew 74 percent in 2015 and reached 3.7 exabytes per month at the end of 2015, which was nearly 4,000 times the one in 2005. Cisco also forecasts that mobile data traffic will surpass 30.6 exabytes per month in 2020. Even though smart devices only represent 36 percent of the total mobile devices and connections, they account for 89 percent of the mobile data traffic. The widespread availability of smart devices equipped with a rich set of built-in sensors has also enabled a new sensing paradigm, mobile crowd sensing, for collecting and sharing sensing data from surrounding environment. This new sensing paradigm makes the mobile data explosion severer.

The current cellular networks do not have enough capacity to support all of the fast-growing mobile big data from these smart devices. Different offloading solutions (such as WiFi networks [2], [3] or femtocells [4]) have been adopted. According to Cisco [1], fifty-one percent of total mobile data traffic was offloaded onto the fixed network through Wi-Fi or femtocell in 2015, and this is the first time offload traffic exceeded cellular traffic. Recently, offloading cellular traffic through opportunistic device-to-device (D2D) communications [5]–[7] among mobile phones becomes a new and possible solution. Compared with current WiFi or femtocell solutions, this method uses occasional D2D contact opportunities to deliver data rather than the fixed network infrastructure. The major advantage is low cost and easy to deploy. Han et al. [5] study how to select the initial set of mobile users to push the content to all users in the networks via D2D, and their proposed heuristics can improve the delivery efficiency and offload a large fraction of data from the cellular network. Li et al. [6] study the problem of multiple mobile data offloading through D2D among different data subscribers under resource constraints. Zhu et al. [7] study offloading peer to peer traffic among mobile users with D2D relays. In this paper, we focus on offloading data collection for mobile sensing data via D2D relays instead broadcasting traffic from the service provider to all subscriber users (as in [5], [6]) or peer-to-peer traffic between any two users (as in [7]).

In most existing mobile sensing systems [8]–[11], the sensing data is collected via cellular networks with the assumption that the size of sensing data is not large. However, with the new types of multimedia sensing and increasing number of sensing devices, the amount of mobile sensing data grows to a scale that traditional cellular methods may not handle. Wang et al. [12] first consider leveraging the delay-tolerant mechanisms by offloading the data to Bluetooth/WiFi gateways or data-plan users. The major goal for their method is to reduce the energy consumption and data cost of data-plan users. Karaliopoulos et al. [13] consider a joint user recruitment problem for both sensing and data collection, where the data collection is done via D2D communications. They formulate the selection of users as a minimum cost set cover problem and propose greedy heuristics to solve it. However, the solution has large time complexity due to the huge search space over all space-time paths across the network, which makes it not suitable for large-scale data collection. In this paper, we only focus on the data collection phase of mobile data sensing by carefully selecting a few relay nodes to help with data propagation via D2D relays. By doing so, we limit the search space and make our algorithm more efficient. In addition, since we use multiple space-time paths for data collection from the source (in [13] only one space-time path is selected for one source), our method can achieve better delivery ratio too.

In summary, in this paper, we study how to select a small subset of relaying devices so that the data propagation via these D2D relays can achieve certain level of delivery ratio. We formulate the problem as an optimization problem in Section II and propose a simple but efficient solution in Section III. In Section IV, we conduct experiments over a real-life mobile

The work is partially supported by the US National Science Foundation under Grant No. CNS-1319915 and CNS-1343355, and the National Natural Science Foundation of China under Grant No. 61428203 and 61572347.

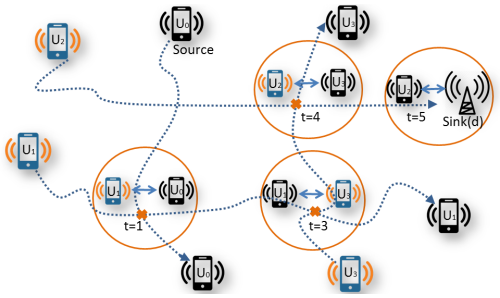


Fig. 1. Example of data delivery via multi-hop D2D communications. Dashed curves are trajectories of devices, a circle represents an encounter between two devices, and the device in black indicates that it has a copy of the data.

trace to confirm the effectiveness of the proposed algorithms.

II. System Model and Problem Statement

A. System Model

We consider the relay node selection problem for data collection. We assume that a mobile user set $User = \{u_1, u_2, ..., u_n\}$, which includes n mobile users who are willing to participate into the delivery of the sensing data. This candidate set is assumed very large given the popularity of smartphones. Each mobile users can visit m different locations, denoted as $Location = \{l_1, l_2, ..., l_m\}$. The whole time period is evenly divided into T sequential time slots, thus time $t \in [1, T]$. Each user has her own visiting pattern over both temporal and spacial domains, i.e., mobile user u_i has her own probability $p(i, j, t)$ to make a visit at location l_j during time slot t. We also assume that these visiting probabilities are independence to the others for each particular mobile users.

To enable device-to-device communications, we assume that two nodes can discover each another and transfer sensing data to each another when they both visit the same location within a particular time slot. For each piece of sensing data, it is generated at a source node s (a mobile device), and needs to be delivered to a sink node d (a mobile device or a static device at certain location). For simplicity, we only focus on the selection of relay nodes for the collection of a single piece of sensing data. However, the method is general enough to handle multiple data pieces and multiple sources/sinks. To enhance the delivery probability, we assume that restricted flooding (i.e., Epidemic [14]) is used within the selected relay nodes. Fig. 1 shows an example of data delivery via multi-hop D2D relays. Note there we assume the data collection is through only device-to-device communications, while in reality, a hybrid solution (combining D2D and direct communication with cellular tower) could be desired.

B. Problem Statement: Relay Node Selection

The key challenging is how to identify a small set of relay nodes from the huge candidate pool in $User$ while guarantee certain level of data delivery. This is different with traditional DTN routing, in which relay nodes are dynamically selected during the routing. We can formally define the relay selection problem as the following optimization problem.

Definition 1: Given the volunteering users $User$ (with their historical call and location traces), and the source s and destination d of the sensing data, *Minimum Relay Problem* is to find a subset $U(s, d)$ of mobile users from $User$ as the relay nodes with the objective to

$$\min |U(s, d)|$$
$$s.t. \;\; p(U(s, d), s, d) \geq \gamma.$$

in which $p(U(s, d), s, d)$ is denoted as the probability that the sensing data can be delivered to its destination sink and γ represents a threshold of the probability.

Similarly, the optimization problem can be defined as another formulation.

Definition 2: Given the volunteering users $User$ (with their historical call and location traces) , and the source s and destination d of the sensing data, *K Relay Problem* is to find a subset $U_{s,d}$ of K mobile user from $User$ as the relay nodes with the objective to

$$\max p(U(s, d), s, d)$$
$$s.t. \;\; |U(s, d)| = K.$$

Note that both versions of the problem are computational challenging, since even with perfect predication of visiting patterns this problem can be reduced to a set cover problem which is NP-hard. Therefore, in this paper, we are looking for efficient heuristics to address this problem.

III. Relay Node Selection for D2D Collection

Recall that we use a flooding/epidemic strategy to deliver the sensing data via multiple hops among selected relay nodes. The selection criteria for relay nodes may rely on how to estimate the delivery probability of a particular group of relay nodes. To achieve this, we first introduce a space-time graph based method, then we propose our greedy based algorithm for relay node selection.

A. Estimation of Delivery Probability via Space-Time Graphs

To capture the evolving characteristics in both spacial and temporal spaces, we adopt the *space-time graph* [15] to model the time-evolving D2D links among selected relay nodes. Let $U(s, d) = \{u_1, \cdots, u_r\}$ is the relay nodes selected for source s and sink d. We can define a space-time graph $\mathcal{G}^{U(s,d)} = (\mathcal{V}, \mathcal{E})$, which is a directed graph defined in both spacial and temporal spaces. Hereafter, we simply use $\mathcal{G}$ to represent $\mathcal{G}^{U(s,d)}$. In $\mathcal{G}$, $T + 1$ layers of nodes are defined and each layer has $r + 2$ nodes (corresponding to $\{u_0 = s, u_1, \cdots, u_r, u_{r+1} = d\}$), thus the whole vertex set $\mathcal{V} = \{u_j^t | j = 0, \cdots, r + 1 \; and \; t = 0, \cdots, T\}$ and there are $(r + 2)(T + 1)$ nodes in total. Fig. 2 illustrates the corresponding space-time graph for the network shown in Fig. 1. Two kinds of links (spacial links and temporal links) are added between consecutive layers in the edge set $\mathcal{E}$. A temporal link $\overrightarrow{u_j^{t-1} u_j^t}$ (those horizontal links in Fig. 2) connects the same node u_j across consecutive $(t-1)$th and tth layers, which represents the node carrying the data in the

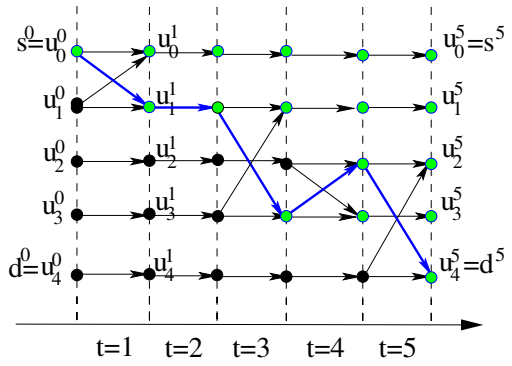


Fig. 2. **Space time graph**: the corresponding space-time graph $\mathcal{G}$ of Fig. 1, where a space-time path from the source s to the sink d is highlighted.

tth time slot. A spacial link $\overrightarrow{u_j^{t-1}u_k^t}$ represents a forwarding possibility from one node u_j to its encountering node u_k in the tth time slot (i.e., u_j encounters u_k in time slot t). By defining the space-time graph $\mathcal{G}$, any communication operation in the time-evolving network can be simulated on this directed graph. E.g., the propagation path in Fig. 1 is highlighted in Fig. 2.

To estimate the delivery probability, we need first define the link probability $p(e)$ of each link $e \in \mathcal{E}$), i.e. the probability of existing such a link. For each temporal link $\overrightarrow{u_j^{t-1}u_j^t}$, its link probability is set to 1 since a node can always hold the data. For a spacial link $\overrightarrow{u_j^{t-1}u_k^t}$, its link probability is calculated as follows.

$$p(\overrightarrow{u_j^{t-1}u_k^t}) = (1 - \prod_{i=1}^{m}(1 - p(j,i,t)p(k,i,t))) \cdot r(\overrightarrow{u_j^{t-1}u_k^t}),$$

where $1 - \prod_{i=1}^{m}(1 - p(j,i,t)p(k,i,t))$ is the probability that node u_j and u_k are co-located at any location and $r(\overrightarrow{u_j^{t-1}u_k^t})$ is the link reliability (representing the successful transfer over the encounter). If u_k is a location l_k instead of a mobile user, $p(\overrightarrow{u_j^{t-1}u_k^t}) = p(j,k,t) \cdot r(\overrightarrow{u_j^{t-1}u_k^t})$.

We then define the delivery probability of a space-time graph $\mathcal{G}$ is $p^{\mathcal{G}}(s^0, d^T)$ regarding the source s and destination d. It is the probability that a packet sent from node s over the routing topology $\mathcal{G}$ reaches node d under flooding-based routing. Similar definition is used in [16] as broadcast reliability. To efficiently calculate this delivery probability is not an easy job. Actually, it is known that the computation of such reliability over general graphs is a problem of NP-hard [17]. Fortunately, the nice loop-free property of our space-time graph model allows us to compute the reliability very efficiently with a dynamic programming algorithm [16]. Basically, for any node u_i^t in $\mathcal{G}$, its delivery probability from the source node s can be calculated as follows:

$$p^{\mathcal{G}}(s^0, u_i^t) = 1 - \prod_{\overrightarrow{u_j^{t-1}u_i^t} \in \mathcal{G}} (1 - p^{\mathcal{G}}(s^0, u_j^{t-1})p(\overrightarrow{u_j^{t-1}u_i^t})).$$

Given the structure $\mathcal{G}$ defined by r relay nodes, starting from a source node, the dynamic programming algorithm can compute the delivery ratio of all other nodes within time of $O(rT(\log(rT) + r))$. Notice that the time complexity of DP algorithm is the same with that of Dijkstra's algorithm. Given the relay node set $U(s,d)$ for source s and sink d, we can estimate the delivery probability based on the space-time graph $\mathcal{G}$ as follows $p(U(s,d), s, d) = p^{\mathcal{G}}(s^0, d^T)$.

Algorithm 1 Relay Selection Algorithm

Input: potential user set $User$, call probability $p(i,j,t)$ for each user in $User$, the source s and the sink d.
Output: selected relay nodes $U(s,d)$.

```
U(s,d) = ∅
while G^{U(s,d)} is connected do
    Choose the most active user and add it into U(s,d)
    while |U(s,d)| < K or p(U(s,d),s,d) < γ (for K relay problem or minimum relay problem, respectively) do
        for all u_i ∈ User and ∉ U(s,d) do
            Calculate the improvement of p(U(s,d),s,d) by adding u_i in to U(s,d)
        Select the user u_i with the largest reliability improvement and add it into U(s,d)
return U(s,d)
```

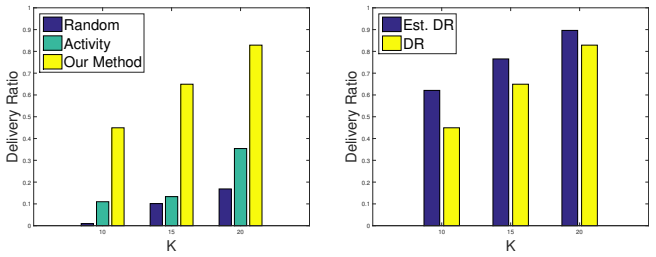
B. Relay Selection Algorithm

Then the relay selection algorithm is quite straightforward. In each step, we greedily select the user u_i which leads to maximal improvement of $p(U(s,d), s, d)$ into $U(s,d)$. Repeat this until either the delivery probability reaches the threshold γ for minimum relay problem or $U(s,d)$ has K users for k relay problem. However, there is still a starting problem, since initially when $U(s,d)$ is empty or just with a few users the space time graph $\mathcal{G}^{U(s,d)}$ may not be connected at all (i.e., $p(U(s,d), s, d) = 0$). In this case, adding any single user may not improve the delivery probability. Therefore, instead of considering improvement of $p(U(s,d), s, d)$, we simply pick the user who is the most active (in term of visited locations). Detailed algorithm is given as Algorithm 1.

IV. Experiments over D4D Dataset

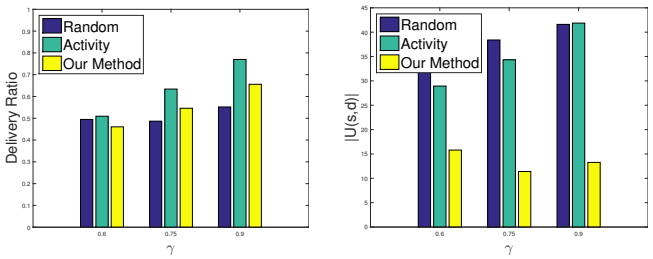
To evaluate the proposed algorithms, we conduct preliminary simulations over a real life cellular data set, D4D dataset [18]. To make comparisons, we also implement two simple heuristics: *random selection* and *activity-based selection*. Random selection randomly chooses a user at each step until the algorithm ends, while activity-based selection greedily chooses a user who is most active (visiting most locations) at each step. In our simulations, we choose the real delivery ratio and the number of selected users as the metrics of measurement.

D4D Dataset and Experiment Settings: The D4D dataset is a data set of large-scale cellular users released by Orange S.A. in 2013, which is based on Call Detail Records of cell phone calls and SMS exchanges among Orange mobile users in Ivory Coast between December 1, 2011 and April 28, 2012. The number of the users is more than $50,000$ for each week. Each record is associated with a cellular tower which provides the service and there are more that $1,000$ cellular towers in this dataset. We select 20 most popular towers, i.e. the towers



(a) delivery ratio (DR) (b) estimated DR vs DR

Fig. 3. Results for K relay problem where $K = 10$, 15 or 20.



(a) delivery ratio (b) number of relay nodes

Fig. 4. Results for minimum relay problem where $\gamma = 0.6$, 0.75 or 0.9.

with largest number of associated records, to implement our simulations. Thus, $m = 20$. We choose a 100 candidate user set as $User$, i.e., $n = 100$. If two users make phone calls associated with same tower at same time, we assume that they are close to each other and could transfer data between them. For simplicity, we set the link reliability as 0.5, i.e., the successful transferring over a pair of nodes is 50% during their encountering. For each data collection task, we randomly select a mobile user as the data source and one location as the sink. For each set of experiments, we test 15 tasks and 100 rounds per tasks. The average performances over $1,500$ rounds are reported.

Experiments on K Relay Problem: In the first set of simulations, we consider the K relay problem. We vary the number of selected relay nodes from 10 to 20. Fig. 3(a) shows the delivery rate achieved by each algorithm. It is clear that our proposed algorithm achieves the highest delivery ratio among the three algorithms when the number of selected relay nodes are the same. In addition, we can find that the delivery ratio of all the three algorithms increase as the number of selected relay nodes increase. This is obvious since more selected relay nodes provide more possible routes for the data to reach the sink node. Fig. 3(b) shows the comparison between expected delivery ratio and real delivery ratio of our proposed algorithm. The real delivery ratio is always lower than the expected one since the estimation is based on the historical records. Although it is not 100 percent accurate, the expected delivery ratio still provides us the guidance to pick the relay nodes.

Experiments on Minimum Relay Problem: In this set of simulations, we evaluate the performance of algorithms over minimum relay problem. Here we vary the delivery ratio threshold γ from 0.6 to 0.9. Fig. 4(a) shows that the delivery ratios of the three algorithms are similar to each others. Recall that all algorithms will continue add new relay nodes until the estimated delivery probability reach the threshold. Since the threshold is the same for the three algorithms, the overall delivery ratios are similar. However, in Fig. 4(b), we can see that the number of selected relay nodes of our algorithm is much fewer than those of the other algorithms when achieving similar level of delivery ratios. This confirms that our proposed algorithm is more efficient than the other two simple heuristics.

V. Conclusion

In this paper, we investigate the feasibility of collecting data packets from mobile devices through device-to-device communications by carefully selecting the subset of relaying devices. We formulate the problem as optimization problems (K relay selection or minimum relay selection) and propose simple greedy algorithms to solve it. The proposed algorithms use historical information to obtain the estimated delivery probability of a given relay set and greedily select the relay node based on this estimated probability. Our experiments over the real-life D4D mobile traces confirm the effectiveness of the proposed algorithms. We leave the study of hybrid data collection scheme which combine D2D and direct communications as our future work.

References

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015-2020 (February 3, 2016). http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.

[2] K. Lee, I, Rhee, J. Lee, S. Chong, and Y. Yi, "Mobile data offloading: how much can WiFi deliver?" in *Proc. of ACM Co-NEXT*, 2010.

[3] S. Dimatteo, P. Hui, B. Han, and V.O.K. Li, "Cellular traffic offloading through WiFi networks," in *Proc. of IEEE MASS*, 2011.

[4] V. Chandrasekhar, J.G. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Communications Magazine*, 46(9):59-67, 2008.

[5] B. Han, et al., "Mobile data offloading through opportunistic communications and social participation," *IEEE Trans. on Mobile Computing*, 11(5):821-834, 2012.

[6] Y. Li, et al., "Multiple mobile data offloading through disruption tolerant networks," *IEEE Trans. on Mobile Computing)*, 13(7):1579–1596, 2014.

[7] Y. Zhu, C. Zhang, and Y. Wang, "Mobile data delivery through opportunistic communications among cellular users: A case study for the D4D challenge," in *Proc. of NetMob*, 2013.

[8] H. Xiong, et al., "EMC^3: Energy-efficient data transfer in mobile crowdsensing under full coverage constraint," *IEEE Trans. on Mobile Computing (TMC)*, 14(7):1355–1368, 2015.

[9] H. Xiong, et al., "Crowdtasker: Maximizing coverage quality in piggyback crowdsensing under budget constraint," in *IEEE Percom*, 2015.

[10] H. Li, T. Li, Y. Wang, "Dynamic participant recruitment of mobile crowd sensing for heterogeneous sensing tasks," in *Proc. of IEEE MASS*, 2015.

[11] H. Li, T. Li, F. Li, et al., "Enhancing participant selection through caching in mobile crowd sensing," in *Proc. of ACM/IEEE IWQoS*, 2016.

[12] L. Wang, et al., "Effsense: energy-efficient and cost-effective data uploading in mobile crowdsensing," in *Proc. of ACM UbiComp*, 2013.

[13] M. Karaliopoulos, et al., "User recruitment for mobile crowdsensing over opportunistic networks," in *Proc. of IEEE INFOCOM*, 2015.

[14] A. Vahdat, et al., "Epidemic routing for partially connected ad hoc networks," Technical Report CS-200006, Duke Univ., Tech. Rep., 2000.

[15] S. Merugu, M. Ammar, and E. Zegura, "Routing in space and time in networks with predictable mobility," Tech. Rep. GIT-CC-04-07, 2004.

[16] F. Li, S. Chen, M. Huang, Z. Yin, C. Zhang, and Y. Wang, "Reliable topology design in time-evolving delay-tolerant networks with unreliable links," *IEEE Trans. on Mobile Computing*, 14(6):1301–1314, 2015.

[17] A. Agrawal and R. E. Barlow, "A survey of network reliability and domination theory," *Operations Research*, 32:478–492, 1984.

[18] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki, "Data for development: The D4D challenge on mobile phone data," in *arXiv.1210.0137v2*, 2013.