

Waze-Inspired Spectrum Discovery via Smartphone Sensing Data Fusion

Sen Lin, Junshan Zhang and Lei Ying

School of EECE, Arizona State University, Tempe, AZ 85287

Abstract—We study Waze-inspired spectrum discovery, where the cloud collects the spectrum sensing results from many smartphones and predicts location-specific spectrum availability based on information fusion. Observe that with limited sensing capability, each smartphone can sense only a limited number of channels; and further, the more channels each smartphone senses, the less accurate the sensing results would be. To develop a comprehensive understanding, we cast the spectrum discovery problem as a matrix recovery problem, which is different from the classical matrix completion problem, in the sense that it suffices to determine only part of the matrix entries in the matrix recovery formulation. It is shown that the widely-used similarity-based collaborative filtering method would not work well because it requires each smartphone to sense too many channels. With this motivation, we propose a location-aided smartphone data fusion method and show that the channel numbers each smartphone needs to sense could be dramatically reduced. Moreover, we analyze the partial matrix recovery performance by using the location-aided data fusion method, and numerical results corroborate the intuition that with each smartphone sensing more channels, the recovery performance improves at first but then degrades beyond some point because of the decreasing sensing accuracy.

I. INTRODUCTION

The past decade has witnessed a skyrocketing demand for commercial wireless spectrum. A report by Credit Suisse shows that wireless base stations in the United States were operating at 80% of their maximum capacity during busy periods and that average smartphone data usage tripled in 2011. These sharp increases in mobile traffic are projected to continue in the foreseeable future. To meet the rapidly growing demand, regulatory agencies (e.g., FCC) around the world are actively working on policies and regulations for dynamic spectrum sharing.

In this paper, we propose a Waze-inspired smartphone spectrum sensing, where the cloud [1] gathers spectrum sensing results from smartphones and uses data fusion to provide the location-specific spectrum information in response to users' requests, as illustrated in Fig. 1. Waze is a highly successful GPS-based mobile navigation application that gathers complementary traffic information from its users and provides real-time traffic information based on user-submitted travel times and route details to help users decide/adjust their routes in real-time. A key observation is that thanks to technological advances, the current generation smartphones already have multiple radio frequency (RF) chains for sensing and communication. In [2], the authors designed and implemented a smartphone-based spectrum sensing scheme by leveraging the spectrum sensing capabilities of WiFi cards, which achieved

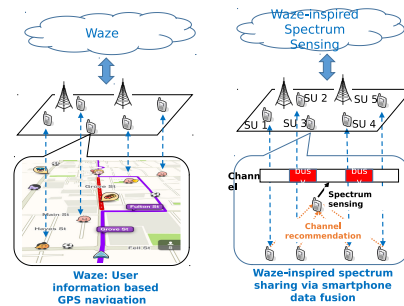


Fig. 1. Waze-inspired spectrum discovery via smartphone sensing data fusion

decent accuracy in determining TV whitespaces. Since one of the primary contributors to the explosive mobile traffic growth is the rapid proliferation of smartphones, a natural question to ask is: “Can future generation smartphones carry out real-time sensing of the communication environment, through built-in spectrum sniffers and shared spectrum usage information from mobile devices that are wirelessly-connected?” With this insight, we shall explore Waze-inspired spectrum discovery via smartphone sensing data fusion. As the number of smartphones is even larger than the worldwide human population, any service enabled by smartphones can likely achieve economics at scale and result in a global reach.

Nevertheless, smartphone sensing suffers from some limitations. (1) *Low accuracy*: compared with high-end spectrum analyzers like ThinkRF analyzer [3], smartphones usually have a smaller sampling rate, resulting in lower accuracy for detecting signal activity. In addition, the sensing accuracy degrades when smartphones sense more channels because of the decrease of sample sizes in each channel given the fixed sensing energy. (2) *Limited observations*: each smartphone often has a restricted view of the spectrum activities due to the limited spectrum sensing capability, and consequently the information about most of the spectrum remains unknown at its location.

The discovery of spectrum availability from reported sensing data can be formulated as a matrix completion problem by completing the observation matrix with smartphones as rows and channels as columns. The cloud must figure out the unknown entries of the matrix from the limited observations of smartphones. When smartphones sense more channels, the performance of matrix completion may not result in more information about spectrum availability, because the sensing accuracy of each individual channel degrades. We will investigate this matrix completion problem and the tradeoff

therein between the sensing accuracy and the number of sensed entries.

One conventional approach to matrix completion is similarity-based collaborative filtering where partial knowledge of neighbors' preference is leveraged to make a prediction about a given user's preference. Neighbors are selected by computing similarity scores. However, our analysis shows this method would not work well in this setting since it would require each smartphone to sense too many channels for matrix completion, thereby beyond the sensing capability of smartphones. To tackle this challenge, we propose a location-aided data fusion method based on another key advantage that smartphones have knowledge of their location information via built-in GPS. Since each smartphone needs one channel to use at each time instant, it suffices to delve into only a partial recovery of the matrix instead of completing the entire observation matrix.

The main contributions in this paper can be summarized as follows.

- *Waze-inspired spectrum discovery*: We propose a Waze-inspired cloud-assisted spectrum sensing model, where each smartphone senses several channels and the cloud makes channel recommendations based on the data fusion of the sensing results from smartphones. In particular, we formulate location-aided spectrum sensing as a matrix recovery problem and use data fusion to determine the spectrum availability.
- *Location-aided smartphone data fusion*: Realizing that one smartphone could provide more reliable sensing of spectrum availability at its very location compared with its neighbors, we apply two different data fusion approaches for the discovery of sensed channels and that of unsensed channels, respectively, in the proposed location-aided method. Theoretical analysis shows the location-aided data fusion method could dramatically reduce the number of channels that smartphones need to sense, superior to the similarity-based collaborative filtering method.
- *Partial matrix recovery*: We characterize the number of recovered entries in the matrix, given the number of channels sensed by each smartphone, by using location-aided smartphone sensing data fusion method. Numerical results validate the effectiveness of our proposed method. As expected, the discovery performance of sensed channels is better than that of unsensed channels thanks to the help of smartphone's own sensing results. Besides, the tradeoff between the sensing accuracy and the number of sensed entries is also depicted in the numerical results, by showing that the number of recovered entries seems to be a quasi-concave function with respect to the number of channels sensed by each smartphone.

The rest of the paper is organized as follows. We introduce the system model in Section II, and present the spectrum discovery methods in Section III. The partial matrix recovery based on location-aided data fusion method in the finite case and the proposed tradeoff problem are studied in Section IV,

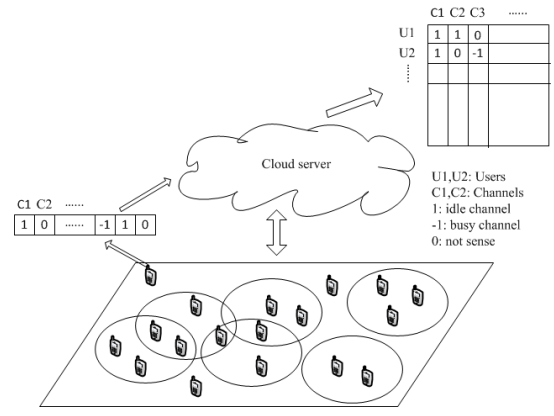


Fig. 2. Cloud-based spectrum discovery via Smartphone sensing data fusion followed by related work in Section V. We conclude in Section VI.

II. SYSTEM MODEL

For ease of exposition, we consider an square area $D \times D$ with N available frequency bands and M smartphones randomly locating in this area. In the spectrum discovery model, each smartphone independently and randomly senses K frequency bands and sends sensing results to the cloud. The cloud gathers all the received results, which can be represented as a matrix where rows represent smartphones and columns represent frequency bands, as shown in Fig. 2. There is no information exchange directly among smartphones.

A. Smartphone Sensing Model

Channel sensing is usually modeled as a binary hypothesis test, where H_0 indicates an idle channel and H_1 indicates a busy channel. Let $r_n(j)$ be the sample collected by the smartphone for channel n , i.e.,

$$r_n(j) = \begin{cases} u_n(j), & H_0 \\ s_n(j) + u_n(j), & H_1 \end{cases}$$

where $j = 1, 2, \dots, J$, J is the number of samples, $u_n(j) \sim \mathcal{CN}(0, \sigma_u^2)$ is the AWGN and $s_n(j)$ is the received user's signal at the smartphone in channel n .

Each smartphone makes decision of channel states based on energy detection. According to [4], the miss detection probability P_m and the false alarm probability P_f can be approximated as

$$P_m = P(H_0|H_1) = 1 - Q \left(\left(\frac{\tau}{\sigma_u^2} - \gamma - 1 \right) \sqrt{\frac{J}{2\gamma + 1}} \right) \quad (1)$$

and

$$P_f = P(H_1|H_0) = Q \left(\left(\frac{\tau}{\sigma_u^2} - 1 \right) \sqrt{J} \right) \quad (2)$$

where τ is the energy decision threshold, γ is the average received user signal power and $Q(\cdot)$ is the Q -function. And the received $\text{SNR} = \gamma / \sigma_u^2$.

Given a fixed sensing energy budget, we assume that each smartphone could at most collect C_u samples in total, such

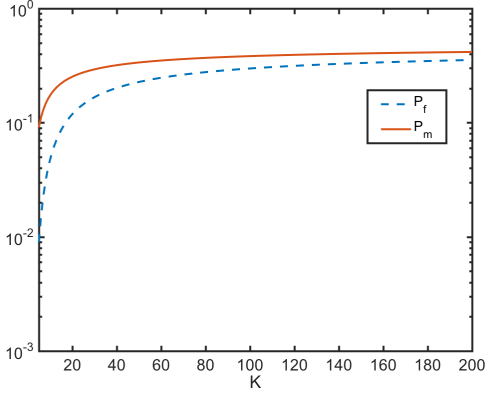


Fig. 3. P_f and P_m under different K , given $C_u = 70000$, $\tau = 1.02$, $\sigma_u^2 = 1$, $\gamma = -15dB$

that on average for each channel $J = \frac{C_u}{K}$ where $C_u > K$. Fig. 3 shows an example of how P_m and P_f change with K .

B. Matrix Representation

For convenience, we define the following matrices:

- *Channel observation matrix*: $E \in R^{M \times N}$, where for smartphone u and frequency band n ,

$$E_{un} = \begin{cases} 1, & \text{if the sensing result for } n \text{ of } u \text{ is busy} \\ 0, & \text{if smartphone } u \text{ did not sense } n \\ -1, & \text{if the sensing result for } n \text{ of } u \text{ is idle} \end{cases}$$

Here binary sensing result is used to reduce the communication cost between the cloud and smartphones.

- *Underlying channel state matrix*: $R \in R^{M \times N}$, where for smartphone u and frequency band n ,

$$R_{un} = \begin{cases} 1, & \text{if frequency band } n \text{ is busy for } u \\ -1, & \text{if frequency band } n \text{ is idle for } u \end{cases}$$

And we assume $P(R_{un} = 1) = P(R_{un} = -1) = \frac{1}{2}$. Similarly, define $\hat{R} \in R^{M \times N}$ as the predicted channel state matrix.

Suppose each smartphone independently and randomly senses each channel with probability $P_s = \frac{K}{N}$, the total number of channels sensed by each smartphone is independently and identically distributed Binomial random variable with parameter $n = N$ and $p = P_s$. It is shown in [5] that the probability of ‘failure’ under the uniform model is bounded by 2 times the probability of ‘failure’ under the Binomial model. Hence, we can restrict our attention to the Binomial model because it admits a simpler analysis than the uniform model thanks to the independence of sampling.

C. Spatial Correlation Model

Since smartphones close to one another may experience similar spectrum availabilities and this similarity would decrease with their distance [6], we assume there exists an effective distance d_0 so that for smartphone u and v

$$P(R_{un} = R_{vn}) = f(d_{uv})$$

where

$$f(d_{uv}) = \begin{cases} e^{-\frac{d_{uv} \log 2}{d_0}} & \text{if } d_{uv} < d_0 \\ \frac{1}{2} & \text{if } d_{uv} \geq d_0 \end{cases}$$

and d_{uv} is the distance between smartphone u and v . We say u and v are ‘‘good’’ neighbors if $d_{uv} < d_0$; otherwise they are ‘‘bad’’ neighbors because one could not provide any useful information about channel states to the other. Define P_d as the probability that u and v are good neighbors, then $P_d = \frac{\pi d_0^2}{D^2}$ considering M smartphones randomly locate in a given area. Similarly,

$$P_\epsilon = Pr(f(d_{uv}) \geq \frac{1}{2} + \epsilon) = P_d \left(\frac{\log \frac{1}{0.5+\epsilon}}{\log 2} \right)^2$$

where $\epsilon \in [0, 0.5]$. Note that the spatial correlation assumption is used for analyzing the spectrum discovery algorithms.

From the spectrum availability discovery perspective, the cloud needs to recover the underlying channel state matrix R from the incomplete observation matrix E . Here we say channel n for smartphone u can be accurately discovered if $P(\hat{R}_{un} \neq R_{un}) \leq P_t$ where P_t is some predefined threshold.

III. WAZED-INSPIRED SPECTRUM DISCOVERY VIA SMARTPHONE DATA FUSION

In what follows, we show that utilizing the location information of smartphones could decrease the order of K from $\omega(\sqrt{N \log N})$ to $\omega(\log N)$ for matrix completion, by comparing the asymptotic performance of traditional collaborative filtering method and that of the proposed location-aided data fusion method. We assume $M = \Theta(N)$ in this section.

A. Similarity-based Collaborative Filtering Method

Similarity-based collaborative filtering is prevalent in practical recommendation systems because it has competitive performance on real datasets with low computation complexity, and is easy to implement. Usually there are two steps in the approach: Step I) to calculate similarity scores between two users to select similar neighbors, and Step II) to produce a prediction of the active user by taking weighted average of all the ratings of selected neighbors on a certain item.

In this work, the similarity score between smartphone u and v is defined as [7]:

$$s_{uv} = \frac{\sum_{n=1}^N \mathbf{1}(E_{un} = E_{vn}, E_{un} \neq 0, E_{vn} \neq 0)}{\sum_{n=1}^N \mathbf{1}(E_{un} \neq 0, E_{vn} \neq 0)}$$

and we apply the following similarity-based method [8] where majority voting is selected as the data fusion method for analysis simplicity to make prediction \hat{R}_{un} of the underlying truth R_{un} :

- *Repeat* For every smartphone u ,
- (i) *Neighbor Selection*: compute similarity scores s_{uv} with all smartphones, select the neighbors with $s_{uv} \geq s_{th} = \frac{1}{2}[1 + (P_m - P_f)^2 + \epsilon(1 - P_f - P_m)^2]$, put them in set \mathcal{N}_u ;

- (ii) *Information Fusion*: for every channel n , define \mathcal{S}_{un} as the set of neighbors in \mathcal{N}_u who sensed frequency band n , compute

$$\hat{R}_{un} = 2 \times \mathbf{1} \left(\frac{1}{|\mathcal{S}_{un}|} \sum_{v \in \mathcal{S}_{un}} E_{vn} \geq 0 \right) - 1$$

• *end*

The value of similarity score threshold s_{th} is chosen such that good neighbors for each smartphone will be selected with high probability. The prediction result is same with the sensing result of majority of selected neighbors. Based on this method, we have the following result on the asymptotic condition in terms of K for matrix completion:

Proposition 1.

- (a) If $K [2\epsilon(1 - P_f - P_m) - |P_m - P_f|]^2 = \omega(\sqrt{N \log N})$, the similarity based collaborative filtering algorithm can recover the underlying channel state matrix R with probability at least $1 - \delta$ for any $\delta > 0$.
- (b) Good neighbors could not be differentiated from bad neighbors if $K = o(\sqrt{N})$.

The proof of Proposition 1 is in the Appendix. It is clear that $K = \omega(\sqrt{N \log N})$ since $[2\epsilon(1 - P_f - P_m) - |P_m - P_f|]^2 \leq 1$. Intuitively, part (a) indicates that the channel state matrix can be recovered if the number of correctly sensed channels for each smartphone is of order $\omega(\sqrt{N \log N})$. When N is large, however, it is not realistic for a smartphone to sense so many frequency bands with decent accuracy. Further, part (b) states that the order of K could not be decreased significantly for similarity-based method to work. One of the underlying rationale in the similarity-based method is due to the fact that good neighbors provide useful information for learning channel states whereas bad neighbors only increase the variance in the data fusion. When $K = o(\sqrt{N})$, the similarity score could not correctly reflect the correlation between two smartphones because there are not enough co-sensed frequency bands for them, resulting in the failure of similarity-based collaborative filtering algorithm.

B. Location-aided Smartphone Data Fusion Method

From the above, it is clear that similarity-based collaborative filtering method is not suitable for Waze-inspired spectrum sensing since it requires each smartphone to sense at least \sqrt{N} channels for distinguishing good neighbors from bad neighbors. However, it is known that smartphones have a lot of capabilities besides spectrum sensing. One key observation is that smartphones can know their location from the built-in GPS. To fully utilize the potential benefit of location awareness, we assume that smartphones also share their location with the cloud besides the spectrum information, based on which we devise a location-aided data fusion approach for matrix recovery. In this case the cloud would differentiate good neighbors from bad neighbors according to locations.

Specifically, we introduce two different data fusion methods for sensed channels and unsensed channels, respectively. It is

clear that the sensing results of each smartphone itself are more accurate than those of its good neighbors in general. Therefore, for sensed channels weighted combination is applied to smartphone's own sensing results and the information from its good neighbors, whereas for unsensed channels we still use majority voting among its good neighbors.

The main idea of location-aided smartphone sensing data fusion method is outlined as follows:

- *Repeat* For every smartphone u ,
- (i) *Neighbor Selection*: the cloud selects neighbor v as a good neighbor if $d_{uv} \leq d_\epsilon$, put v in set \mathcal{N}_u ;
- (ii) *Information Fusion*: define \mathcal{S}_{un} as the set of neighbors in \mathcal{N}_u who sensed frequency band n ,
 - for frequency bands with $E_{un} \neq 0$, compute

$$\hat{R}_{un} = 2 \times \mathbf{1} \left(\lambda E_{un} + (1 - \lambda) \frac{1}{|\mathcal{S}_{un}|} \sum_{v \in \mathcal{S}_{un}} E_{vn} \geq 0 \right) - 1$$

- for frequency bands with $E_{un} = 0$, compute

$$\hat{R}_{un} = 2 \times \mathbf{1} \left(\frac{1}{|\mathcal{S}_{un}|} \sum_{v \in \mathcal{S}_{un}} E_{vn} \geq 0 \right) - 1$$

• *end*

where $d_\epsilon = d_0 \frac{\log \frac{1}{0.5+\epsilon}}{\log 2}$, $\epsilon \in [0, 0.5]$ and $\lambda = \frac{1}{1+\alpha|\mathcal{S}_{un}|}$,

$$\alpha = \frac{2|P_f - P_m + 2\epsilon(1 - P_f - P_m)|}{\{1 - (P_f - P_m)^2 - 4\epsilon(1 - P_f - P_m)[P_f - P_m + \epsilon(1 - P_f - P_m)]\} \log \frac{1 - P_m}{1 - P_m}}$$

Here $\lambda \in [0, 1]$ is chosen to minimize the discovery error probability. It can be seen that when $|\mathcal{S}_{un}|$ is larger (which means more good neighbors sense channel n), λ is smaller. This is intuitive since the cloud should trust the good neighbors more when $|\mathcal{S}_{un}|$ is larger considering the low possibility that most of the good neighbors make a mistake simultaneously.

Based on this method, we have the following result:

Proposition 2. If $K [2\epsilon(1 - P_f - P_m) - |P_m - P_f|]^2 = \omega(\log N)$, the location-based information fusion algorithm can recover the underlying channel state matrix R with probability at least $1 - \delta$ for any $\delta > 0$.

The proof of Proposition 2 is in the Appendix. Since the cloud could identify good neighbors based on the location information, Proposition 2 indicates that with correctly sensed channels of order $\omega(\log N)$ for each smartphone, there are adequate good neighbors providing enough useful information in the data fusion, such that the matrix could be recovered.

C. Performance Comparison

To better understand the performance improvement for location-aided data fusion, we first consider the case $P_m = P_f$, and it is worth noting that the general cases can be translated to this special case by selecting suitable energy detection threshold. Since the sensing capability of smartphones plays an important role in determining the order of K , we investigate

$$I_{th}^1 = \frac{2[1 - (P_f - P_m)^2 + 4\epsilon(1 - P_m - P_f)(P_m - P_f) - 4\epsilon^2(1 - P_m - P_f)^2]}{[P_f - P_m + 2\epsilon(1 - P_f - P_m)]^2} \log \frac{1}{2P_t} \quad (3)$$

$$I_{th}^2 = \frac{2[1 - (P_f - P_m)^2 + 4\epsilon(1 - P_m - P_f)(P_f - P_m) - 4\epsilon^2(1 - P_m - P_f)^2]}{[P_m - P_f + 2\epsilon(1 - P_f - P_m)]^2} \log \frac{1}{2P_t} \quad (4)$$

$$I_{th}^3 = \frac{[1 - (P_f - P_m)^2 + 4\epsilon(1 - P_m - P_f)(P_f - P_m - 4\epsilon(1 - P_m - P_f))]}{2[P_m - P_f + 2\epsilon(1 - P_f - P_m)]^2} \left(2 \log \frac{\sqrt{P_m(1 - P_m)}}{P_t} + \sqrt{4 \log^2 \frac{\sqrt{P_m(1 - P_m)}}{P_t} - (1 - 2P_m)^2} \right) \quad (5)$$

$$I_{th}^4 = \frac{[1 - (P_f - P_m)^2 + 4\epsilon(1 - P_m - P_f)(P_m - P_f - 4\epsilon(1 - P_m - P_f))]}{2[P_f - P_m + 2\epsilon(1 - P_f - P_m)]^2} \left(2 \log \frac{\sqrt{P_f(1 - P_f)}}{P_t} + \sqrt{4 \log^2 \frac{\sqrt{P_f(1 - P_f)}}{P_t} - (1 - 2P_f)^2} \right) \quad (6)$$

TABLE I
PERFORMANCE COMPARISON UNDER DIFFERENT C_u

	similarity-based data fusion method	location-aided data fusion method
$C_u = o(\log N)$	Fails	Fails
$C_u = o(\sqrt{N \log N})$	Fails	If $K = \omega(\log N)$, works
$C_u = \omega(\sqrt{N \log N})$	If $K = \omega(\sqrt{N \log N})$, works	If $K = \omega(\log N)$, works

K under different cases of smartphone sensing ability in terms of total sampling number C_u .

From equation (2) and $K < C_u$, we have that

$$K(1 - 2P_f)^2 = K \left(1 - 2Q \left(\left(\frac{\tau}{\sigma_u^2} - 1 \right) \sqrt{\frac{C_u}{K}} \right) \right)^2 = \Theta(K)$$

Based on Proposition 1 and Proposition 2, Table 1 gives an overview of the performance of the similarity-based method and the location-aided method under different C_u and K .

The similarity-based data fusion method works only if smartphones have great sensing capability and $K = \omega(\sqrt{N \log N})$, whereas the location-aided data fusion method works even if smartphones have weak sensing capability and only requires smartphones to sense $\omega(\log N)$ channels.

IV. PARTIAL MATRIX RECOVERY IN THE FINITE CASE

In this section, we study the partial matrix recovery in the finite case, by using the location-aided smartphone sensing data fusion method. We first characterize the number of discovered channels for a given K , and then demonstrate the tradeoff between the sensing accuracy and the number of sensed entries. Moreover, numerical results show that the number of discovered channels seems to be a quasi-concave function of K , which pin-points to the tradeoff impact on the partial matrix recovery performance.

A. Partial Matrix Recovery

Motivated by the fact that each smartphone only needs one channel at each time instant, we next characterize the exact relation between K and the number of discovered channels, in the finite case.

For convenience, we define

$$G_u = M(N - K)g(I_{th}^u, K)$$

$$H_u = \sqrt{M(N - K)[1 - g(I_{th}^u, K)] \log N}$$

$$G_s = MKg(I_{th}^s, K)$$

$$H_s = \sqrt{MK[1 - g(I_{th}^s, K)] \log N}$$

and

$$g(t, K) = Q \left(\frac{t - M \frac{K}{N} P_\epsilon}{\sqrt{M \frac{K}{N} P_\epsilon (1 - \frac{K}{N} P_\epsilon)}} \right)$$

where $I_{th}^u = \max(I_{th}^1, I_{th}^2)$, $I_{th}^s = \max(I_{th}^3, I_{th}^4)$ and the values of I_{th}^1 , I_{th}^2 , I_{th}^3 and I_{th}^4 are shown in the equations (3)-(6). We have the following proposition:

Proposition 3. *By using the location-aided smartphone sensing data fusion method,*

- at least $L_u = G_u - 2H_u - \log N$ unsensed channels can be discovered;
- at least $L_s = G_s - 2H_s - \log N$ sensed channels can be discovered.

The proof of Proposition 3 can be found in the Appendix. Proposition 3 characterizes the number of unsensed channels and sensed channels that can be discovered, respectively, for a given K . Let $I_{un} = |S_{un}|$ denote the number of neighbors who sense channel n among all selected neighbors of smartphone u . We have the following remarks:

- First we investigate the minimum value of I_{un} , required to determine the availability of a channel, regardless of whether it has been sensed by smartphone u or not. Observe that more good neighbors will provide more useful information for channel discovery. Denote I_{th}^u and I_{th}^s as the threshold values of I_{un} for unsensed and sensed channels, respectively. It is clear from equations (3)-(6) that both I_{th}^u and I_{th}^s are determined by the sensing accuracy in terms of P_m and P_f . Specifically, the more accurate the sensing results are, the less good neighbors spectrum discovery would need. In addition, $I_{th}^u > I_{th}^s$, because the presence of smartphone's own sensing results alleviates the need of good neighbors for the discovery of sensed channels.
- Next, L_u and L_s characterize the number of unsensed entries with $I_{un} \geq I_{th}^u$ and that of sensed entries with $I_{un} \geq I_{th}^s$, respectively. Taking L_u as an example,

intuitively G_u points to the average number of unsensed entries with $I_{un} \geq I_{th}^u$, while H_u represents the variation around the average number.

Note that L_u and L_s are nonnegative since they represent the number of unsensed channels and sensed channels that could be discovered, respectively. When $G_u < 2H_u + \log N$ (or $G_s < 2H_s + \log N$), we say $L_u = 0$ (or $L_s = 0$) such that no channels can be discovered.

B. Tradeoff between the Sensing Accuracy and the Number of Sensed Entries

Recall that when K increases, P_m and P_f will increase due to the limited sensing capability of smartphones, and as a result the sensed entries in the observation matrix E become less accurate. Meanwhile, the matrix E will have more sensed entries which could lend more useful information for the matrix recovery. Hence there is clearly a tradeoff between the sensing accuracy and the number of sensed entries.

To quantify the tradeoff in the partial matrix recovery problem, we formulate the following optimization problem:

$$\begin{aligned} \max_{K, \epsilon, \tau} \quad & L = L_u + L_s \\ \text{s.t.} \quad & K \in [1, N], \\ & \epsilon \in [0, 0.5] \end{aligned}$$

To better understand the impact of the sensing accuracy on the recovery performance, we compare the following two cases:

- *The case where P_m and P_f are constant:* This is a general assumption in the literature on matrix completion [7]. As a result, I_{th}^u and I_{th}^s are constant. As K increases, the matrix can be recovered, i.e., $L = MN$, because $g(I_{th}^u, K)$ and $g(I_{th}^s, K)$ would approach to 1.
- *The case where P_m and P_f increase with K :* This is the case in our model. Both I_{th}^u and I_{th}^s increase with K accordingly. In the low SNR region, $g(I_{th}^u, K)$ and $g(I_{th}^s, K)$ would approach to 0 when K is large enough, and hence no entries could be recovered, i.e., $L = 0$. It can be shown that $g(I_{th}^u, K)$ and $g(I_{th}^s, K)$ could be approximated by quasi-concave functions of K .

In a nutshell, the fact that sensing accuracy decreases with K makes it impossible for the matrix recovery under large K in the low SNR region.

Delving into the specific relation between L and K , it can be seen that L is a monotonically decreasing function of I_{th}^u and I_{th}^s for a given K , and hence the optimal τ should be selected when I_{th}^u and I_{th}^s are minimized. From equations (3)-(6), both I_{th}^u and I_{th}^s will increase when $P_m \neq P_f$. The numerical results indicate that τ is optimal when $P_m = P_f$. We can first simplify L based on this observation, and it is nevertheless nontrivial to analyze the convexity of L directly due to the complicated expression. Alternately, in what follows we show that L seems to be a quasi-concave function with respect to K in the low SNR region by numerical results, resulted by the tradeoff between the sensing accuracy and the number of sensed entries, under the assumption that M and N are reasonably large.

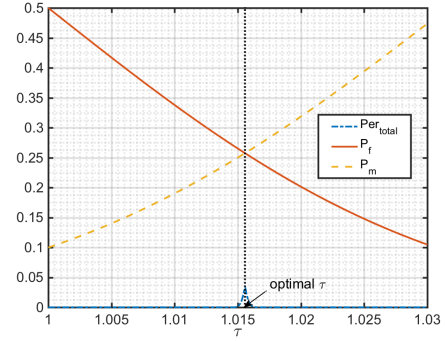


Fig. 4. Impact of energy detection threshold τ on total spectrum discovery percentage, given $\epsilon = 0.16$, $K = 40$ and $C_u = 70000$.

C. Numerical Results

To evaluate the spectrum discovery performance, we define the following metrics to represent the spectrum discovery percentage for unsensed channels and sensed channels, respectively:

$$\begin{aligned} Per(K) &= \begin{cases} \frac{L_u}{M(N-K)} & \text{for unsensed channels} \\ \frac{L_s}{MK} & \text{for sensed channels} \end{cases} \\ Per_{total}(K) &= \frac{L_u + L_s}{MN} \end{aligned}$$

We consider a large area with enough smartphones and channels, by setting $M = 50000$ and $N = 10000$. P_d is chosen as 0.5 to ensure enough neighbors of each smartphone. We further consider the low SNR case where $\gamma = -15dB$, $\sigma_u^2 = 1$ and the corresponding discovery threshold $P_t = 10^{-1}$. Besides, we assume that each smartphone could collect up to $C_u = 70000$ samples. The matrix partial recovery performance depends on τ , K and ϵ . We investigate the impact of those parameters sequentially and show one example of optimal point for the optimization problem.

From Proposition 3, it is clear that the impact of P_f on L is symmetric with that of P_m , while P_m and P_f depend on the selection of τ . Fig. 4 shows the impact of τ on all channel discovery performance Per_{total} . When $\tau = 1.0155$, $P_f = P_m$ and Per_{total} achieves the maximum value. This is because we consider the discovery of busy and idle channels simultaneously, and larger P_m (or P_f) leads to adverse effect on the discovery of busy (or idle) channels. In the following analysis, we choose $\tau = 1.0155$.

The discovery percentages for unsensed channels and sensed channels under different K are illustrated in Fig. 5, respectively. It can be seen that $Per(K)$ first increases with K then continually decreases beyond the optimal point. Additionally, the discovery performance of sensed channels outperforms that of unsensed channels where only neighbors' information is considered, corroborating the advantage of smartphone's own sensing results. Fig. 6 shows the changing tendency of the number of discovered channels. In particular, L_u , L_s and L all seem to be quasi-concave functions with respect to K . This validates the tradeoff effect: When K is small, P_m and P_f are relatively small, such that the increase of sensed entries will

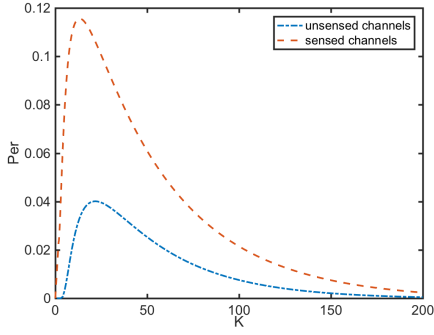


Fig. 5. Impact of K on spectrum discovery percentages given $\tau = 1.0155$, $C_u = 70000$ and $\epsilon = 0.16$.

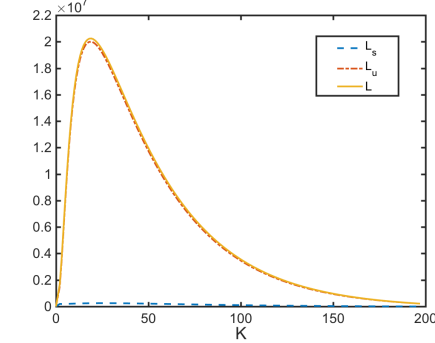


Fig. 6. Impact of K on number of discovered channels given $\tau = 1.0155$, $C_u = 70000$ and $\epsilon = 0.16$.

benefit more than the negative effect caused by the increase of P_m and P_f ; when K grows, the negative effect of increasing P_m and P_f becomes more pronounced than the positive effect, leading to the degradation of the overall spectrum discovery performance. Note that L_s is much smaller than L considering that there are far more unsensed channels than sensed channels in the model. L is maximized when $K \approx 25$. More specifically, around 400 channels can be discovered for each smartphone on average although each smartphone just senses 25 channels with large error probability, which is adequate for a smartphone to select idle channels for data transmission. This demonstrates that the location-aided data fusion method serves as a satisfying solution for the spectrum discovery based on the sensing data of smartphones with limited sensing capability.

The selection of ϵ also affects the matrix partial recovery performance since it would impact the number of good neighbors selected and the spatial correlation. Fig. 7 illustrates the discovery performance under different ϵ . It can be seen that Per_{total} also seems to be a quasi-concave function of ϵ . Note that when ϵ is larger, the neighbors selected will have larger correlation with the smartphone under consideration, while the number of good neighbors selected will decrease.

Moreover, we consider the impact of K and ϵ together on the discovery performance. The optimization contour plot of Per_{total} is shown in Fig. 8, where the optimal point is near $(0.17, 20)$ and the optimal value of L is around 2×10^7 . In this case, more than 400 channels can be discovered for each smartphone when each smartphone only sensed 20 channels.

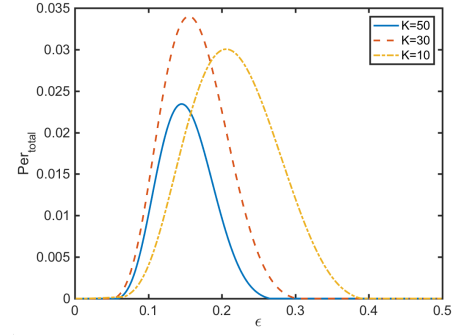


Fig. 7. Impact of ϵ on total spectrum discovery percentage given $K = 10, 30, 50$, $\tau = 1.0155$ and $C_u = 70000$.

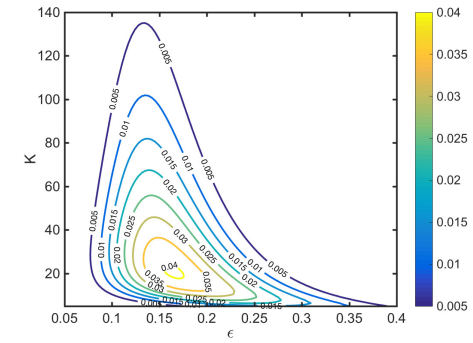


Fig. 8. 2D Optimization contour plot given $\tau = 1.0155$ and $C_u = 70000$.

V. RELATED WORK

Cooperative spectrum sensing: There are a lot of literatures on cooperative spectrum sensing in cognitive radio [9] [10], most of which focus on finding the optimal cooperation methods and reducing cooperation overhead. [11] formulated a non-transferable coalitional spectrum sensing game, wherein each secondary user strategically joins a coalition to maximize its channel detection probability. [12] used linear prediction to aggregate observations from different secondary users to estimate the channel availability probabilities. [13] applied compressive sensing technique to reduce the sensing and feedback overhead. But none of them considered the random sensing model and derived the performance guarantee in smartphone sensing-based spectrum discovery problem.

Collaborative filtering: Collaborative filtering (CF) is popular to solve matrix recovery problem in item recommendation [14], but only a few works considered theoretical analysis for similarity-based CF methods. [15] derived theoretical performance guarantee of a similarity-based algorithm in asymptotic case under clustered user-item model and all users are assumed to rate every item with same probability. Based on this, [7] considered the heterogeneous case where both information-sparse and information-rich users exist in the model. However, all these work assumed users and items are clustered without overlapping while in our model no explicit clusters are formulated. Moreover, to the best of our knowledge, our work is the first study to consider partial matrix recovery in the finite case and the tradeoff between the sensing accuracy and the number of sensed entries.

VI. CONCLUSION

In this work, we propose Waze-inspired spectrum sensing via smartphone data fusion to discover the location-specific multi-channel spectrum availability. We formulate the spectrum discovery problem as a matrix recovery problem and apply location-aided smartphone sensing data fusion method by showing that utilizing location information could greatly decrease the order of K from $\omega(\sqrt{N \log N})$ to $\omega(\log N)$ compared with similarity-based collaborative filtering method. Further, we analyze the performance of the proposed location-aided data fusion method in terms of the partial recovery of the matrix and evaluate the tradeoff between the sensing accuracy and the number of sensed entries.

VII. ACKNOWLEDGEMENT

This work is supported partially by National Science Foundation grants ECCS-1547294 and ECCS-1408409.

REFERENCES

- [1] X.Chen and J.Huang, "Database-assisted distributed spectrum sharing," *IEEE Journal on Selected Areas in Communications*, 2013.
- [2] T. Zhang, A. Patro, N. Leng, and S. Banerjee, "A wireless spectrum analyzer in your pocket," in *HotMobile'15*, 2015.
- [3] ThinkRF, "The thinkrf wireless signals intelligence platform." <http://www.thinkrf.com>.
- [4] Y. C. Liang, Y. Zeng, E. C. Y. Peh, and A. T. Hoang, "Sensing-throughput tradeoff for cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 4, pp. 1326–1337, 2008.
- [5] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, 2006.
- [6] M. Wellens, J. Riihijarvi, M. Gordziel, and P. Mahonen, "Spatial statistics of spectrum usage: From measurements to spectrum models," in *2009 IEEE International Conference on Communications*, 2009.
- [7] K. Zhu, R. Wu, L. Ying, and R. Srikant, "Collaborative filtering with information-rich and information-sparse entities," *Mach Learn*, vol. 97, pp. 177–203, 2014.
- [8] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, "An algorithmic framework for performing collaborative filtering," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999.
- [9] C.Sun, W.Zhang, and K.B.Letaief, "Cluster-based cooperative spectrum sensing in cognitive radio systems," in *2007 IEEE International Conference on Communications*, 2007.
- [10] L. Wang, J. Wang, G. Ding, F. Song, and Q. Wu, "A survey of cluster-based cooperative spectrum sensing in cognitive radio networks," in *Cross Strait Quad-Regional Radio Science and Wireless Technology Conference (CSQRWC), 2011*, 2011.
- [11] W. Saad, Z.Han, M.Debbah, A.Hjorungnes, and T.Basar, "Coalitional games for distributed collaborative spectrum sensing in cognitive radio networks," in *IEEE INFOCOM 2009*, pp. 2114–2122, 2009.
- [12] H. Li, "Learning the spectrum via collaborative filtering in cognitive radio networks," in *New Frontiers in Dynamic Spectrum, 2010 IEEE Symposium on*, 2010.
- [13] J. J. Meng, W. Yin, H. Li, E. Hossain, and Z. Han, "Collaborative spectrum sensing from sparse observations in cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, 2011.
- [14] X. Yang, Y. Guo, Y. Liu, and H. Steck, "A survey of collaborative filtering based social recommender systems," *Computer Communications*, vol. 41, pp. 1–20, March 2014.
- [15] K. Barman and O. Dabeer, "Analysis of a collaborative filter based on popularity amongst neighbors," *IEEE Trans. Inf. Theory*, 2012.
- [16] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.
- [17] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: A survey," *Internet Mathematics*, vol. 3, no. 1, pp. 79–127, 2006.

VIII. APPENDIX

Due to space limitations, we present only a brief sketch of the proofs here.

A. Proof of Proposition 1

There are two steps in the similarity-based collaborative filtering method: Step I) to distinguish good neighbors from bad neighbors by similarity score and Step II) to make predictions based on the information of selected good neighbors. To prove part (a) in Proposition 1, we have the following lemma.

Lemma 1. *If $K[2\epsilon(1 - P_f - P_m) - |P_m - P_f|]^2 = \omega(\sqrt{N \log N})$, all the neighbors selected by the similarity-based collaborative filtering method are good neighbors with probability at least $1 - cN^{-1}$ where c is some constant.*

Lemma 2. *If the following conditions hold:*

- 1) *All good neighbors could be selected correctly;*
- 2) *$K[2\epsilon(1 - P_f - P_m) - |P_m - P_f|]^2 = \omega(\log N)$;*

matrix R can be recovered by the similarity-based method with probability at least $1 - \delta$ for any $\delta > 0$.

Lemma 2 indicates the conditions such that majority voting among good neighbors could make good predictions. By combining Lemma 1 and 2, we prove (a) in Proposition 1.

For the second part, by using Lemma 7 in [15], it can be shown that $P(s_{uv} \leq s_{th} | K = o(\sqrt{N})) \geq C_1$ if u and v are good neighbors, and $P(s_{uv} \geq s_{th} | K = o(\sqrt{N})) \geq C_2$ if u and v are bad neighbors, where C_1 and C_2 are some positive constants. Many bad neighbors are selected so that this similarity-based method failed to distinguish good neighbors from bad neighbors.

B. Proof of Proposition 2

In the location-aided data fusion method, all good neighbors can be distinguished by the cloud according to locations. λ is derived by minimizing the spectrum discovery error probability. The proof of Proposition 2 is based on Lemma 2 and the generalized Chernoff bound [17].

C. Proof of Proposition 3

The proof includes two steps: Step I) to figure out the minimum value of I_{un} required to determine the availability of a channel, and Step II) to characterize the number of entries with I_{un} above the minimum value, as explained previously.

In the first step, since we consider that M is reasonably large and the random sensing model such that the data fusion is a sum of I_{un} i.i.d Bernoulli random variables, the Central Limit Theorem can be used to derive the minimum value of I_{un} , I_{th}^u for unsensed channels and I_{th}^s for sensed channels, given that $P(\hat{R}_{un} \neq R_{un}) \leq P_t$.

Next, suppose we want to discover L_u channels among $M(N - K)$ unsensed channels, we can obtain L_u by ensuring that $I_{un}^{(L_u)}$ is larger than I_{th}^u with high probability, where $I_{un}^{(L_u)}$ denotes the L_u -th largest one among $I_{un}^1, \dots, I_{un}^{M(N-K)}$ and I_{un}^i represents the number of good neighbors who sensed i -th channel among all $M(N - K)$ unsensed channels. L_s can also be derived following the same idea.