Kui Ren Xue Liu Weifa Liang
Ming Xu Xiaohua Jia Kai Xing (Eds.)

LNCS 7992

Wireless Algorithms, Systems, and Applications

8th International Conference, WASA 2013
Zhangjiajie, China, August 2013
Proceedings

Springer

Kui Ren Xue Liu Weifa Liang
Ming Xu Xiaohua Jia Kai Xing (Eds.)

Wireless Algorithms, Systems, and Applications

8th International Conference, WASA 2013
Zhangjiajie, China, August 7-10, 2013
Proceedings

Springer

Volume Editors

Kui Ren
University at Buffalo, NY, USA
E-mail: kuiren@buffalo.edu

Xue Liu
McGill University, Montreal, QC, Canada
E-mail: xueliu@cs.mcgill.ca

Weifa Liang
Australia National University, Canberra, ACT, Australia
E-mail: wliang@cs.anu.edu.au

Ming Xu
National University of Defense Technology, Changsha, China
E-mail: xuming@nudt.edu.cn

Xiaohua Jia
City University of Hong Kong, Kowloon, China
E-mail: csjia@cityu.edu.hk

Kai Xing
University of Science and Technology of China, Hefei, Anhui, China
E-mail: kxing@ustc.edu.cn

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-39700-4 e-ISBN 978-3-642-39701-1
DOI 10.1007/978-3-642-39701-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013943126

CR Subject Classification (1998): F.2, C.2, C.4, H.4

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Over the past few years, wireless communications and networks have enjoyed tremendous growth, driven by 3G/4G cellular technologies, wide deployment of WiFi access points, and proliferation of smart personal mobile devices. At the same time, end users are accustomed to bandwidth-hungry applications such as online video streaming, online gaming, emails with multimedia attachment, etc. Emergence of multimedia networking requires next-generation wireless networks to provide for not only basic Internet access but also quality of service guarantee, seamless roaming across heterogeneous networks. Scalable solutions are crucial for handling large amounts of mobile users; and they give rise to new challenges, for both industry and academia, in resource allocation and scheduling, mobility management, distributed algorithms, cooperative networking and dynamic spectrum sharing, security and privacy, scalable and energy-efficient network protocols.

The annual International Conference on Wireless Algorithms, Systems, and Applications (WASA) provides a forum for theoreticians, system and application designers, protocol developers and practitioners to exchange ideas, share new findings, and discuss challenging issues for the current and next-generation wireless networks. Past WASA conferences were held in Xi'an (2006), Chicago (2007), Dallas (2008), Boston (2009), Beijing (2010), Chengdu (2011), and Yellow Mountains (2012). The 8th WASA conference, took place in Zhang Jiajie during August 7–10, 2013. The conference accepted 43 high-quality papers, 25 of which were selected from the 65 open submissions plus 18 invited papers. All the papers went through peer reviews by the Technical Program Committee. The conference received 80 full submissions. Each submission was rigorously reviewed by at least three Program Committee members.

We thank all the authors for submitting their papers to the conference. We also thank all the members of the Technical Program Committee and external referees for their help in completing the reviewing process under the tight time constraints. We especially thank Special Session Chairs, Dr. Xinwe Fu, Dr. Jianwei Huang, Dr. Zhi Sun, Dr. Xiuzhen Cheng, and Dr. Honggang Wang for inviting high-quality papers. We are grateful to the members of the Steering Committee and Organizing Committee for their involvement throughout the process. WASA 2013 was an event of true teamwork. Finally, many other people have contributed to the success of WASA 2013, whose names cannot be listed here due to space limitation. However, we owe them our gratitude.

August 2013

Kui Ren
Xue Liu
Weifa Liang

Organization

WASA 2013 was organized by the National University of Defense Technology, China, in cooperation with NSFC.

Steering Committee

Xiuzhen Cheng	The George Washington University, USA, Co-chair
Peng-Jun Wan	Illinois Institute of Technology, USA, Co-chair
Ness Shroff	The Ohio State University, USA
Ty Znati	University of Pittsburgh, USA
Wei Zhao	University of Macau, China
Jiannong Cao	Hong Kong Polytechnic University, Hong Kong, SAR China

Executive Committee

Honorary General Chair

Xingming Zhou	National University of Defense Technology, China

General Co-chairs

Ming Xu	National University of Defense Technology, China
Xiaohua Jia	Hong Kong City University, Hong Kong, SAR China

TPC Co-chairs

Kui Ren	SUNY Buffalo, USA
Xue Liu	McGill University, Canada
Weifa Liang	Australia National University, Australia

Local Organization Chair

Yongjun Wang	National University of Defense Technology, China

Publicity Co-chairs

Limin Sun	Institute of Software Chinese Academy of Sciences, China

Xinbing Wang	Shanghai Jiaotong University, China
Laurence T. Yang	St. Francis Xavier University, Canada
Zhi Sun	SUNY Buffalo, USA

Publication Co-chairs

Kai Xing	University of Science and Technology of China, China
Wei Peng	National University of Defense Technology, China

Panel Chair

Junshan Zhang	Arizona State University, USA

Registration Chair

Jiahao Wang	University of Electronic Science and Technology, China

Finance Chair

Dong Wang	Hunan University, China

Technical Program Committee

Jianguo Yao	Shanghai Jiao Tong University, China
Thomas Nolte	MRTC/Malardalen University, Sweden
Ming Li	Utah State University, USA
Nei Kato	Tohoku University, Japan
Ning Cao	Google, USA
Rui Chu	National University of Defense Technology, China
Xiaofu Wu	Nanjing Institute of Communications Engineering, China
Xiuzhen Cheng	The George Washington University, USA
Zhi Sun	SUNY Buffalo, USA
Kai Xing	University of Science and Technology of China
Minming Li	City University of Hong Kong, SAR China
Hamed Mohsenian-Rad	University of California, Riverside, USA
Chonggang Wang	InterDigital Communications, USA
Habib M. Ammari	University of Michigan-Dearborn, USA
Yu Wang	University of North Carolina at Charlotte, USA
Huan Li	Beihang University, China
Qian Wang	Illinois Institute of Technology, USA
Jian Tang	Syracuse University, USA
Hongwei Du	Harbin Institute of Technology, China

Shaoliang Peng	National University of Defense Technology, China
Jiming Chen	Zhejiang University, China
Haojin Zhu	Shanghai Jiaotong University, China
Liang Zhou	Nanjing University of Posts and Telecommunications, China
Tommaso Melodia	State University of New York at Buffalo, USA
Qun Li	College of William and Mary, USA
Chi Zhang	University of Science and Technology of China
Qian Zhang	Hong Kong University of Science and Technology, SAR China
Wonjun Lee	Korea University, South Korea
Weiyi Zhang	AT&T Research Lab, USA
Lei Ying	Arizona State University, USA
Jianwei Niu	Beijing University of Aeronautics and Astronautics, China
Hongbo Jiang	Huazhong University of Science and Technology, China
Jiang Linda Xie	University of North Carolina at Charlotte, USA
Zhou Su	Waseda University, Japan
Xiaoxia Huang	Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
Zhu Han	University of Houston, USA
Pengjun Wan	Illinois Institute of Technology, USA
Shuguang Cui	Texas A&M University, USA
Kui Wu	University of Victoria, Canada
Jianping Wang	City University of Hong Kong, SAR China
Su Gang	Huazhong University of Science and Technology, China
Jing Liu	Shanghai Jiaotong University, China
Kyung-Joon Park	Daegu Gyeongbuk Institute of Science and Technology, South Korea
Benyuan Liu	University of Massachusetts Lowell, USA
Dong Xuan	Ohio State University, USA
Lisong Xu	University of Nebraska, Lincoln, USA
Hwangnam Kim	Korea University, South Korea
Sung-Soo Lim	Kookmin University, South Korea
Chen Tian	Huazhong University of Science and Technology, China
Pan Li	Mississippi State University, USA
Rong Zheng	McMaster University, Canada
Mohammad Husain	California State University, Pomona, USA
Song Ci	University of Nebraska-Lincoln, USA

Table of Contents

Improving Particle Filter with Better Proposal Distribution for Nonlinear Filtering Problems

Fasheng Wang[1,2], Xucheng Li[1,2], and Mingyu Lu[1,*]

[1]School of Information Science and Technology, Dalian Maritime University,
No.1 Linghai Road, 116026 Dalian, China
[2]Dept. of Compt. Sci. and Techn., Dalian Neusoft Univ. of Inform.,
No.8 Software Park Road, 116023 Dalian, China
{fswang,lumingyu}@dlmu.edu.cn, lixucheng@neusoft.edu.cn

Abstract. Designing better proposal distributions can greatly affect the performance of the particle filters, which has been extensively studied in the literature. In this paper, we propose to design better proposal distribution using a new version of unscented Kalman filter- the iterated unscented Kalman filter (IUKF). The IUKF makes use of both statistical and analytical linearization techniques in different steps of the filtering process, which makes it a better candidate for designing proposal distribution in particle filter framework. Each particle is updated using an iterative manner. Through this process, the algorithm can make better use of the current observation for state estimation. To evaluate the performance of the proposed particle filter, we use a synthetic model and a real-world model for the experiments. The experimental results have shown that the proposed algorithm outperforms the alternatives.

Keywords: Particle Filter, Proposal Distribution, Iterated Unscented Kalman filter, Option Pricing.

1 Introduction

Particle filter has grown to be a standard solution for nonlinear filtering problems. It has been applied with great success to a variety of fields including computer vision [1,2], signal processing [3], target tracking [4], financial options pricing [5], etc.

The basic idea underlying this algorithm is to represent the posterior probability density function (PDF) $p(x_k|y_{1:k})$ using a set of weighted random samples (or particles) which are drawn from a carefully designed proposal distribution. Doucet [6] has proved that the optimal proposal distribution that can minimize the variance of the particle weights is: $q(x_k|x_{0:k-1}, y_{1:k}) = p(x_k|x_{0:k-1}, y_{1:k})$. But it cannot be used efficiently in practice due to the fact that we cannot know exactly the PDF before the particles are drawn. Researchers have been working on this topic and some suboptimal proposal distributions were proposed. Among of them, the most famous one is the transition prior $p(x_k|x_{k-1})$ which is proposed

[*] Corresponding author.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 1–12, 2013.

by Gordon [7]. The transition prior based particle filter (generic particle filter) is easy to implement and has been widely used in solving real-world problems, especially in visual tracking field. Isard [1] first used the algorithm (CONDENSATION) in visual tracking. However, the transition prior is not always effective due to not incorporating the latest observations. In order to incorporate the recent observations to improve the estimation accuracy of particle filter, Doucet [8] introduced the local linearization based approximation technique, which is a popular proposal distribution devising method. This method relies on the first order Taylor expansions of the system models. The extend Kalman filter (EKF) is often used within the particle filtering framework to generate proposal distribution, which is call the extended Kalman particle filter (EKPF). But the linearization step can introduce inaccuracies leading to divergence of the EKPF.

Li Liang-qun [9] proposed to use the iterated extend Kalman filter (IEKF) as the proposal distribution yielding the iterated extended Kalman particle filter (IEKPF). IEKF can reduce the linearization errors by iteratively updating the state estimation , which makes the IEKF a better candidate for designing proposal distributions. The unscented Kalman filter (UKF) is also an excellent candidate in particle filter framework. It is a kind of linear regression Kalman filter and can yield elegant performance without linearization steps as required by the EKF which make it much more suitable for proposal distribution. Merwe [10] used the UKF as the proposal distribution within the particle filter framework resulting in unscented particle filter (UPF). The UPF has been successfully applied to visual tracking, neural networking training, etc., which shows improved performance compared to other alternative algorithms.

Wenyan Guo [11] combined the conventional iterated Kalman filter (IKF) with the square root UKF to generate proposal distributions designing an improved UPF which showed improved performance compared to the UPF. Fasheng Wang [12] proposed to use a hybrid Kalman filter (HKF) as the proposal distribution giving a hybrid Kalman particle filter (HKPF). The HKF is directly combination of the EKF and the UKF. The system states are firstly updated using the EKF equations obtaining preliminary estimations. Then, the preliminary estimations are used as input for the UKF to obtain the final state estimations. The states are updated again using the UKF equations. The idea of the HKF is a bit like the IEKF, and both of them make better use of the current observations. The HKPF can obtain improved estimation accuracy compared to the IEKPF and UPF. However, for nonlinear systems with strong nonlinearity, both of the HKPF and IEKPF can fail to reach satisfactory accuracies due to the drawbacks of the EKF.

In this paper, we propose to use the iterated UKF (IUKF), as the proposal distribution in the particle filter framework. The IUKF [13] has shown efficiency in nonlinear filtering problems and can achieve better performance than IEKF, UKF and EKF. It used the statistical linearization for the prediction step, and the analytical linearization for the update step of filtering. Thus, it is a much better candidate for designing proposal distribution in the particle filter framework.

The rest of this paper is organized as follows. In section 2, we briefly reviewed the generic particle filter algorithm. Section 3 gives the details of the proposed particle filter. Experiments and analysis are given in section 4. The last section draws the conclusion.

2 Particle Filter

Consider the following nonlinear dynamic models:

$$x_{k+1} = f(x_k) + w_k \tag{1}$$

$$y_k = h(x_k) + v_k \tag{2}$$

where x_k is the state vector, w_k is the process noise caused by disturbances and modeling errors. y_k is the observation vector, and v_k is the measurement noise. $f()$ and $h()$ are the state transition function and the measurement function, respectively.

The aim of particle filtering algorithm is to approximate the PDF $p(x_k|y_{1:k})$ using a particle set $\{x_k^i, w_k^i\}_{i=1}^N$,

$$p(x_k|y_{1:k}) = \sum_{i=1}^{N} w_k^i \delta(x_k - x_k^i) \tag{3}$$

All the particles are drawn from a properly designed proposal distribution $q(x_k|x_{0:k-1}, y_{1:k})$. Under first-order Markovian assumption, the particle weight is recursively updated using the following equation:

$$w_k^i = w_{k-1}^i \frac{p(y_k|x_k^i)p(x_k^i|x_{k-1}^i)}{q(x_k^i|x_{k-1}^i, y_k)} \tag{4}$$

Details of the derivation can be found in [14]. The standard particle filter algorithm is summarized as in Algorithm 1.

3 The Iterated Unscented Particle Filter

3.1 Iterated Unscented Kalman Filter

The IUKF is a straightforward application of the scaled unscented transformation which uses a set of carefully chosen weighted sigma points to approximate the state distribution. Suppose the dimension of x_k is L, and state mean and covariance estimations are $\hat{x}_{k|k}$ and $P_{k|k}$, respectively. The sigma points are selected according to the following equations:

$$\chi_0 = \hat{x}_{k|k} \tag{5}$$

$$\chi_i = \hat{x}_{k|k} + \left(\sqrt{(L+\lambda)P_{k|k}}\right)^{(i)}, i = 1, 2, ..., L \tag{6}$$

Algorithm 1. Standard Particle Filter

```
Initialization: $k$=0,
  –For $i$=1,...,$N$,
    – Draw $x_0^i$ from the prior density $p(x_0)$.
  –ENDFor
For $k$=1,2,...
  –For $i = 1, 2, ..., N$,
    – Draw a new particle $x_k^i \sim q(xi_k|x_{k-1}^i, y_k)$
    – Assign the particle a weight $w_k^i$ according to (4).
  –ENDFor
  –For $i = 1, 2, ..., N$,
    – Normalize the weights: $w_k^i = \frac{w_k^i}{\sum_{j=1:N} w_k^j}$.
  –ENDFor
  –Resampling:
    – Multiply or suppress the particles $\{x_k^i, w_k^i\}_{i=1}^N$ with high/low weights $w_k^i$, respectively, to obtain $N$ random samples approximately distributed according to $p(x_k^i|y_{1:k})$.
    – Assign equal weight $1/N$ to all of the resampled particles.
    – Output: The particle set used for approximate the PDF.
ENDFor
```

$$\chi_i = \hat{x}_{k|k} - \left(\sqrt{(L+\lambda)P_{k|k}}\right)^{(i)}, i = L+1, ..., 2L \tag{7}$$

$$W_m^0 = \frac{\lambda}{L+\lambda} \tag{8}$$

$$W_c^i = W_m^i = \frac{1}{2(L+\lambda)}, i = 1, 2, ..., 2L \tag{9}$$

$$W_c^0 = \frac{\lambda}{L+\lambda} + (1 - \alpha^2 + \beta) \tag{10}$$

where $\lambda = \alpha^2(L+\kappa) - L$ is a scaling parameter such that $\lambda + L \neq 0$. $\left(\sqrt{(L+\lambda)P_{k|k}}\right)^{(i)}$ is the i_{th} column or row of the matrix square root of $(L+\lambda)P_{k|k}$, α decides the diffusion of the sigma points around $\hat{x}_{k|k}$ and often set to be a small positive value, κ is a secondary scaling parameter and set to be 0, β is used to incorporate the prior knowledge of the state distribution. W_m^i is the weight for calculating the mean estimate of the sigma points, while W_c^i for covariance estimate.

The sigma points are propagated through the nonlinear model to yield a set of transformed sigma points.

$$\chi_{k+1|k}^i = f(\chi_{k|k}^i) \tag{11}$$

Then we predict the estimated state mean as follow,

$$\hat{x}_{k+1|k} = \sum_{i=0}^{2L} W_{m,k}^i \chi_{k+1|k}^i \tag{12}$$

and the predicted covariance is computed as follow

$$P_{k+1|k} = \sum_{i=0}^{2L} W_{c,k}^{i} \left[\chi_{k+1|k}^{i} - \hat{x}_{k+1|k}\right] \left[\chi_{k+1|k}^{i} - \hat{x}_{k+1|k}\right]^{T} \tag{13}$$

After the current observation y_k is obtained, the predicted estimations are updated. The mean estimate is updated in an iterative manner.

According to Bayesian theorem, $p(x_k|y_{1:k})$ can be written as:

$$p(x_k|y_{1:k}) = p(x_k|y_{1:k-1}, y_k) \tag{14}$$

$$= \frac{1}{m} p(y_k|x_k, y_{1:k-1}) p(x_k|y_{k-1}) \tag{15}$$

$$= \frac{1}{m} p(y_k|x_k) p(x_k|y_{k-1}) \tag{16}$$

where m is a normalizing constant:

$$m = \int p(y_k|x_k) p(x_k|y_{1:k-1}) dx_k \tag{17}$$

We assume that both the predicted state and the measurement noise are Gaussian distributed, so $p(x_k|y_{1:k})$ in (14) is also Gaussian. Finding the maximum a posteriori (MAP) estimation is equivalent to maximizing the logarithm, which is also equivalent to minimizing the following equation:

$$f(x_k) = \frac{1}{2} \left(x_k - \hat{x}_{k|k-1}\right)^T P_{k|k-1}^{-1} \left(x_k - \hat{x}_{k|k-1}\right) + \frac{1}{2} \left(y_k - h(x_k)\right)^T R_k^{-1} \left(y_k - h(x_k)\right) \tag{18}$$

We use the Newton-Raphson iteration beginning from $\bar{x}_0 = \hat{x}_{k|k-1}$. At the j_{th} iteration, $\bar{x}_{j-1}$ is already obtained in the previous iteration. We expand $f(x)$ to second order Taylor series as follow:

$$f(x) = f(\bar{x}_{j-1}) + (x - \bar{x}_{j-1})^T \frac{\partial f(\bar{x}_{j-1})}{\partial x} + \frac{1}{2} (x - \bar{x}_{j-1})^T \frac{\partial^2 f(\bar{x}_{j-1})}{\partial x^2} (x - \bar{x}_{j-1}) \tag{19}$$

where $\frac{\partial f}{\partial x}$ is the gradient and $\frac{\partial^2 f}{\partial x^2}$ is the Hessian of $f(x)$. The minimum estimate of the approximation $\bar{x}_j$ can be computed by equating the gradient of the approximation to be zero. By computing the differentiation of (17) and equating it to be zero, we can obtain the following equation:

$$\bar{x}_j = \bar{x}_{j-1} - \left(\frac{\partial^2 f(\bar{x}_{j-1})}{\partial x^2}\right)^{-1} \frac{\partial f(\bar{x}_{j-1})}{\partial x} \tag{20}$$

where

$$\frac{\partial f(\bar{x}_{l-1})}{\partial x} = P_{k|k-1}^{-1} (\bar{x}_{l-1} - \hat{x}_{k|k-1}) - H_j^T R_k^{-1} (y_k - h(\bar{x}_{l-1})) \tag{21}$$

$$\frac{\partial^2 f(\bar{x}_{l-1})}{\partial x} = P_{k|k-1}^{-1} + H_j^T R_k^{-1} H_j \tag{22}$$

can be obtained according to (18), and $H_j = H(\bar{x}_{j-1})$ is the Jacobian matrix of $h(x)$ evaluated at $\bar{x}_{j-1}$. Thus, the iteration equation can be obtained as follow:

$$\bar{x}_j = \bar{x}_{j-1} - \left(P_{k|k-1}^{-1} - H_j^T R_k^{-1} H_j\right)^{-1} K_{k|k-1} \tag{23}$$

$$K_{k|k-1} = \left[P_{k|k-1}^{-1}(\bar{x}_{j-1} - \hat{x}_{k|k-1}) - H_j^T R_k^{-1}(y_k - h(\bar{x}_{j-1}))\right] \tag{24}$$

The iteration number is usually set to be a fixed value J. The final state estimation at time step k is $\hat{x}_{k|k} = \bar{x}_J$. The covariance estimation is computed using:

$$P_{k|k} = \left(P_{k|k-1}^{-1} + H_j^T R_k^{-1} H_j\right)^{-1} \tag{25}$$

3.2 Iterated Unscented Particle Filter

The IUKF obtains the MAP estimate by iteration in which the current observations are made full use to calculate the final estimations. Thus, the IUKF can achieve more accurate estimations than the UKF, which makes it a better candidate for proposal distribution in particle filter framework. The new filter that results from the IUKF proposal distribution is called the iterated unscented particle filter (IUPF). The IUPF algorithm can be summarized as in Algorithm 2.

3.3 Converge Analysis

We assume that $\mathbf{B}(\mathbb{R}^n)$ is the space of bounded, *Borel* measurable functions on $\mathbb{R}^n$. Let's define $\|f\| \triangleq sup|f(x)|_{x\in\mathbb{R}^n}$. According to theorem 2 in [15], we obtain the following theorem.

Theorem 1. *If the importance weight in (4) is upper bounded for any (x_{k-1}, y_k), then, for all $k \geq 0$, there exists c_k independent of N such that for any $f_k \in \mathbf{B}(\mathbb{R}^n)$,*

$$E\left[\left(\frac{1}{N}\sum_{i=1}^{N} f_k(x_{0:k}^i) - \int f_k(x_{0:k})p(dx_{0:k}|y_{1:k})\right)^2\right] \leq c_k \frac{\|f_k\|^2}{N} \tag{26}$$

This expectation corresponds to the randomness introduced by the particle filter algorithm. This convergence result indicates that under very loose assumptions, the IUPF is sure to converge and that the convergence rate of the method is independent of the dimension of the state-space.

4 Experiments

In this section, we will compare the performance of the IUPF to that of the other particle filters on two estimation problems. The generic PF is based on [7], EKPF and UPF are based on [10], IEKPF is based on [9], and HKPF is based on [12].

Algorithm 2. Iterated Unscented Particle Filter

```
Initialization: $k$=0,
–For $i$=1,...,$N$,
  – Draw $x_0^i$ from the prior density $p(x_0)$ and set:
  – $\hat{x}_0 = E(x_0^i)$
  – $P_0^i = E\left[\left(x_0^i - \hat{x}_0^i\right)\left(x_0^i - \hat{x}_0^i\right)^T\right]$
–ENDFor
For $k$=1,2,...
(a) Importance Sampling step:
–For $i = 1, 2, ..., N$,
  – Update the particles with the IUKF
    —- Calculate the sigma points:
    —- $\chi_{k|k-1}^i = \hat{x}_{k-1|k-1}^i \pm \left(\sqrt{(L+\lambda)P_{k-1|k-1}^i}\right)^{(i)}$
  – Time update:
    —- Let $\bar{x}_0 = \hat{x}_{k-1|k-1}^i$
    —- For $j = 1, 2, ..., J$,
      —— $\bar{x}_j = \bar{x}_{j-1} - \left(P_{k|k-1}^{i,-1} - H_j^T R_k^{-1} H_j\right)^{-1} \left[P_{k|k-1}^{i,-1}(\bar{x}_{j-1} - \hat{x}_{k|k-1}) - H_j^T R_k^{-1}(y_k - h(\bar{x}_{j-1}))\right]$
    —- ENDFor
    —- Let $\hat{x}_{k|k}^i = \bar{x}_J$, and $P_{k|k}^i = \left(P_{k|k-1}^{i,-1} + H_j^T R_k^{-1} H_j\right)^{-1}$
  – Generate proposal distribution $N(\hat{x}_{k|k}^i, P_{k|k}^i)$, and draw a new particle $x_k^i$ from
    the proposal distribution.
  – Set $x_{0:k}^i = (x_{0:k-1}^i, x_k^i)$ and $P_{0:k}^i = (P_{0:k-1}^i, P_{k|k}^i)$.
–ENDFor
–Calculate the particle weights and normalize them.
(b) Resampling as in Algorithm 1.
(c) Output: The particle set used for approximate the PDF.
ENDFor
```

4.1 Synthetic Experiment

In this experiment, we adopt the following system models taken from [10]:

$$x_k = 1 + \sin[0.04\pi(k-1)] + 0.5x_{k-1} + w_{k-1} \tag{27}$$

$$y_k = \begin{cases} 0.2x_k^2 + v_k, & k < 30 \\ 0.5x_k - 2v_k, & k > 30 \end{cases} \tag{28}$$

The system noise w_k is assumed to be a Gamma random variable $\zeta_a(3, 2)$, and the measurement noise is modeled as a Gaussian distribution $N(0, 0.00001)$. The purpose of the algorithms is to estimate the state sequence x_k for $k = 1, 2, ..., 60$, given the observations y_k. The experiment is repeated 100 times with random reinitialization for each run. All the algorithms use 200 particles. The parameters for UPF and IUPF are set to be $\alpha = 1$, $\beta = 0$, and $\kappa = 2$.

Figure 1 compares the estimates generated from a single run of the different filters, which shows that the IUPF can estimate the system states accurately close to the true values. Table 1 shows the performance of different filters which summarizes the means and variances of the mean-square-error (MSE) of the

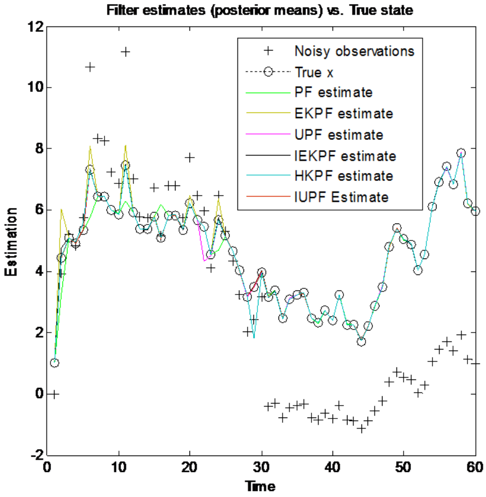


Fig. 1. Plot of estimates generated by different algorithms

Table 1. The means and variances of the MSE

Algorithm	MSE mean	MSE variance
Generic PF	0.55483	0.04853
EKPF	0.32239	0.021157
UPF	0.11159	0.011498
IEKPF	0.044169	0.0018387
HKPF	0.039041	0.0017837
IUPF	***0.016826***	***0.00063118***

state estimates. The mean of the MSE generated by IUPF is about 0.017, while the variance is about 0.00063, which are much less than the other filters. The superior performance of the IUPF is clearly evident.

4.2 Pricing Financial Options

In this experiment, we compare the performance of different filters through a realistic financial options pricing problem. The task of this experiment is to predict the option prices using different filters. We use the famous Black-Scholes partial differential equation as follow [10]:

$$\frac{\partial f_0}{\partial t} + rS\frac{\partial f_o}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 f_o}{\partial^2 S} = r f_o \tag{29}$$

where f_o is the current value of an option, S is the current value of underlying cash product, σ is the volatility of the cash product, and r is the risk-free interest rate. We model the option prices as follow [10,12]:

$$C_p = SN_c(d_1) - Xe^{-rt_m}N_c(d_2) \tag{30}$$

where C_p is the call option price, P_p is the put option price, X is the strike price, t_m is the time to maturity, $N_c()$ is the cumulative normal distribution, and d_1 and d_2 are given as follows:

$$d_1 = \frac{ln(S/X) + (r + \sigma^2/2)t_m}{\sigma\sqrt{t_m}} \tag{31}$$

$$d_2 = d_1 - \sigma\sqrt{t_m} \tag{32}$$

We treat r and σ are the hidden states, C_p and P_p are the output observations. The experiment data is taken from [10] which are five pairs of call and option contracts on the British FTSE 100 index from Feb. 1994 to Dec. 1994.

In this experiment, we use 100 particles and the experiment is repeated 10 times. Figure 2 plots the one-step-ahead predictions on the call and put options prices. The predictions obtained by the IUPF are very close to the true values. The difference between the predictions is much more obvious if the figure is zoomed on. In Table 2, we compare the one-step-ahead normalized square errors (NSE) obtained with each filter on a pair of options with strike price 2925. The NSEs are defined as follows:

$$NSE_{C_p} = \sqrt{\sum_k (C_p^{(k)} - \hat{C}_p^{(k)})^2} \tag{33}$$

$$NSE_{P_p} = \sqrt{\sum_k (P_p^{(k)} - \hat{P}_p^{(k)})^2} \tag{34}$$

where $\hat{C}_p^{(k)}$ and $\hat{P}_p^{(k)}$ are the one-step-ahead predictions of the call and put option prices at the k_{th} day. The NSEs are only measured over the past 100 trading

Table 2. One-step-ahead prediction NSE

Option Type	Algorithm	NSE Mean
Call	Generic PF	0.032732
	EKPF	0.0084599
	UPF	0.008196
	IEKPF	0.0056975
	HKPF	0.0039902
	IUPF	***0.00031026***
Put	Generic PF	0.026569
	EKPF	0.0081043
	UPF	0.008047
	IEKPF	0.0055239
	HKPF	0.0040768
	IUPF	***0.00031856***

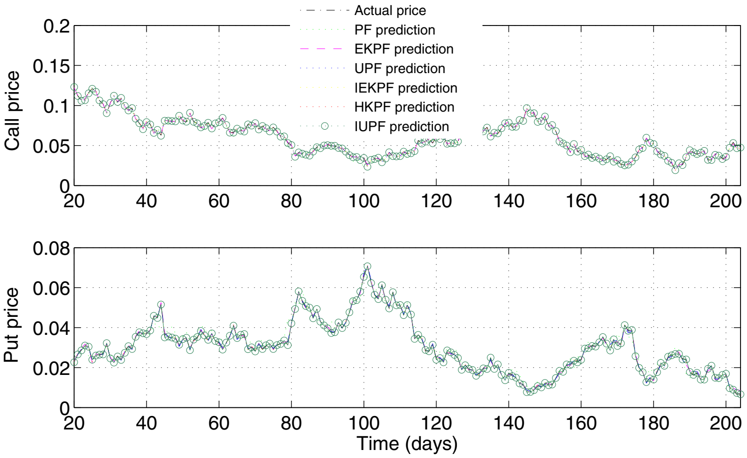


Fig. 2. Predicted prices of the options obtained by different filters

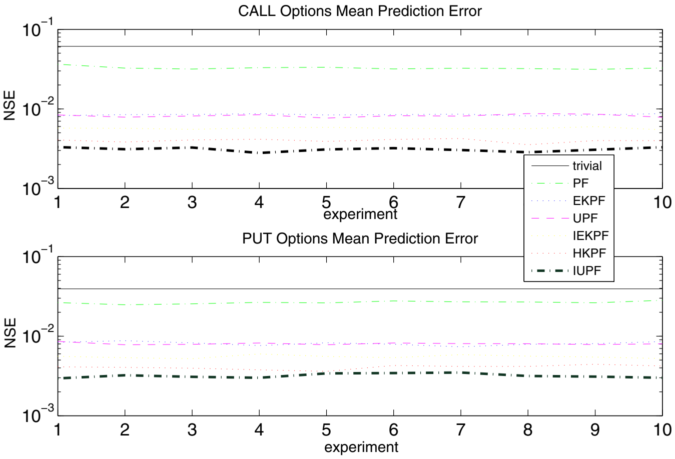


Fig. 3. Plot of NSEs of different filters in each independent run

days in order to allow the algorithms to converge. It shows that the proposed IUPF is superior to the other filters.

Figure 3 shows the NSEs of different filters at each independent run, which shows that the NSEs of the IUPF are continuously less than the other filters. In this figure, the bottom dashed black line is the IUPF prediction NSEs in each run. The top black solid line is the trivial prediction NSE curve. The trivial prediction is obtained by assuming that the price on the following day corresponds to the current price.

5 Conclusion

In this paper, we proposed a new particle filter using the IUKF to generate the proposal distribution. The IUKF can make better use of the current observations through iterations. Thus it is a better candidate for designing proposal distributions than the other alternatives. The proposal distribution generated by IUKF is much closer to the true posterior probability distribution. We demonstrated the superior performance of the IUPF through synthetic simulation and realistic experiment.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (No. 61073133, 61175053, 60973067), the Innovation Group Project of China Education Ministry (No. 2011ZD010), Fundamental research fund of DMU (NO. 2012QN030), Foundation of Scientific Planning Project of Dalian City (No. 2011E15SF100), the Natural Science Foundation of Liaoning Education Ministry (No. L2011241).

References

1. Isard, M., Blake, A.: Condensation-conditional density propagation for visual tracking. Int. J. Comput. Vision. 29, 5–28 (1998)
2. Fasheng, W.: Particle filters for visual tracking. Commun. Comput. Inform. Sci. 152, 107–112 (2011)
3. Doucet, A., Freitas, J., Gordon, N.: Sequential Monte Carlo Methods in Practice. Springer, New York (2001)
4. Fasheng, W., Quan, G., Yuejin, L.: A Kalman Particle Filter for Bearing-only Target Tracking. J. Comput. Inform. Syst. 7, 5628–5635 (2011)
5. Creal, D.: A survey of sequential Monte Carlo methods for economics and finance. Econometric Rev. 31, 245–296 (2012)
6. Doucet, A., Godsill, S., Andrieu, C.: On sequential Monte Carlo sampling methods for Bayesian filtering. J. Stat. Comput. 10, 197–208 (2000)
7. Gordon, N., Salmond, D., Smith, A.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE-Proc. F. 140, 107–113 (1993)
8. Doucet, A.: On sequential simulation-based methods for Bayesian filtering. Technical report, Department of Engineering, Cambridge University (1998)
9. Li, L., Ji, H., Luo, J.: The iterated extended Kalman particle filter. In: International Symposium on Communication and Information Technology, pp. 1172–1175. IEEE Press, New York (2005)
10. Merwe, R., Doucet, A., Freitas, N., Wan, E.: The Unscented Particle Filter. Technical report, Engineering Department, Cambridge University (2000)
11. Guo, W., Han, C., Lei, M.: Improved Unscented Particle Filter for Nonlinear Bayesian Estimation. In: International Conference on Information Fusion, pp. 1–6. IEEE Press, New York (2007)
12. Wang, F., Lin, Y., Zhang, T., Liu, J.: Particle Filter with Hybrid Proposal Distribution for Nonlinear State Estimation. J. Comput. 6, 2491–2501 (2011)

13. Banani, S., Masnadi-Shirazi, A.: A New Version of Unscented Kalman Filter. In: Proc. World Academy of Science, Engineering, and Technology, vol. 20, pp. 192–197 (2007)
14. Arulampalam, M., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for On-line Nonlinear/Non-Gaussian Bayesian Tracking. IEEE T. Signal Process. 50, 174–188 (2002)
15. Crisan, D., Doucet, A.: A Survey of Convergence Results on Particle Filtering Methods for Practitioners. IEEE T. Signal Process. 50, 736–746 (2002)

Performance Evaluation with Control Channel on the Coexistence Scenario of TD-LTE and LTE-FDD

Yinshan Liu[1,2], Xiaofeng Zhong[1], Jing Wang[1],
Yang Lan[3], and Atsushi Harada[3]

[1] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2] Dalian Airforce Communication NCO Academy, Dalian, Liaoning 116600, China
[3] DOCOMO Beijing Communications Laboratories Co., Ltd., Beijing 100190, China
ys-liu10@mails.tsinghua.edu.cn, {zhongxf,wangj}@tsinghua.edu.cn,
{lan,harada}@docomolabs-beijing.com.cn

Abstract. The coexistence performance with control channel when TD-LTE and LTE-FDD systems coexist at the adjacent frequency band within the same geographical area is investigated. A simple mathematical model mapping from control channel performance to data channel throughput is employed to estimate system performance when considering the impact of the adjacent channel interference (ACI) on the control channels. The simulation results indicate that the ACI generated by Base Station (BS) of collocated LTE system significantly degrades the performance of the uplink of the victim LTE system. The link between BSs needs an extra isolation of 83.6 dB in the typical coexistence scenarios.

Keywords: TD-LTE, LTE-FDD, coexistence, ACI, control channel.

1 Introduction

LTE system, which is standardized by 3GPP, has the advantages of high data throughput and low latency, improved system capacity and coverage. The standards cover TDD and FDD modes, which are the TD-LTE and LTE-FDD, and may be deployed by different operators in some countries or regions. GSA (Global mobile Suppliers Association) has confirmed LTE as the fastest developing mobile system technology ever. There have been 145 commercial network launches in 66 countries as of January 8, 2013[1]. GSA anticipates that almost 234 commercial LTE networks will be established in 83 countries by the end of 2013[1]. It is expected that TD-LTE and LTE-FDD will coexist in the foreseeable future, especially in China, and should be operated at the adjacent bands such as 2500-2690 MHz band (band 38 for TD-LTE and band 7 for LTE-FDD), as recommended by ITU. In addition, the technical report TR36.942 released by 3GPP stresses the coexistence issues of two LTE systems and of LTE with other 2G/3G systems [2]. Therefore, the coexistence issue of TD-LTE and LTE-FDD systems is of great research value. Many researches have been done in the area[3–5]. However, these studies only focused on the evaluation of the throughput loss

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 13–24, 2013.

of the victim system from the perspective of the data channel. The control channel (CCH) interfered by the ACI from collocated LTE system has impact on the overall system performance, and this impact has not been investigated so far.

When the victim LTE system is interfered by LTE system nearby, the ACI received by some CCHs is more serious, because they are located near the transmission band of the interfering system, e.g. Physical Uplink Control Channel (PUCCH) and Physical Downlink Control Channel (PDCCH). In addition, the performance of the CCH may limit the overall system throughput, e.g. the error of CCH causes the loss of all resources of the associated data channel since user equipment (UE) cannot identify its resource allocation. Furthermore, the CCH is subject to more constraints, such as limited amounts of both time-frequency and power resource, etc, thus the performance improvement of the CCHs is low, even though they are aided with diversity techniques and resource scheduling.

The main purpose of this investigation is to evaluate the impact of affected CCHs interfered by the ACI on the system throughput. The main contribution of this paper is the development of a general analytical model that proposes a mapping from the CCHs performance to the system throughput, which overcomes the restriction in previous work that only considers the throughput loss of the data channel. The model can *fully* consider the affected LTE system loss by jointly analyzing the control and data channel when another LTE system using adjacent frequency band is activated. This work is simulation based, and is carried out using a system-level LTE simulator. The rest of this paper is organized as follows. The interference scenarios analysis, including coexistence issue and interference analysis, is given in Sect. 2. Section 3 presents a simple mathematical model mapping from CCHs performance to data channels throughput. Section 4 describes the simulation assumptions, parameters and simulation methodology in detail. Section 5 gives simulation results and analysis for system performance with CCH under different control parameters and coexistence deployment scenarios. Finally, Section 6 draws a conclusion.

2 Interference Scenarios Analysis

2.1 Coexistence Issue of LTE-FDD and TD-LTE

LTE is a type of Orthogonal Frequency Division Multiplexing (OFDM) based system whose frame structure, general settings and channel configuration can be found in [7–9]. The LTE physical channels include two data channels and six CCHs. The data channels are Physical Downlink Shared Channel (PDSCH) and Physical Uplink Shared Channel (PUSCH). The downlink (DL) CCHs include Physical Broadcast Channel (PBCH), PDCCH, Physical Control Format Indicator Channel (PCFICH) and Physical Hybrid Indicator Channel (PHICH). The uplink (UL) CCHs include PUCCH and Physical Random Access Channel (PRACH). LTE-FDD/TD-LTE system which acts as victim is affected by the interference, which comes from TD-LTE /LTE-FDD system that acts as aggressor. Due to different uplink-downlink configurations between TD-LTE and

LTE-FDD, there are four types of interference scenarios for the coexistence of both systems:

- TDD/FDD DL to FDD/TDD DL
- TDD/FDD DL to FDD/TDD UL
- TDD/FDD UL to FDD/TDD DL
- TDD/FDD UL to FDD/TDD UL

2.2 Interference Analysis

Since the OFDM technology is employed in LTE system, we consider that there is no internal interference between orthogonal subcarriers.

Co-channel Interference. For Co-Channel Interference (CCI), owing to orthogonal subcarriers in LTE systems, there is no interference between UEs with different resource unit (e.g. resource block, control channel elements) for CCHs and data channel. Only the UEs that use the same resource unit in different sector would be calculated for CCI [3]. The total received CCI is expressed as:

$$I_{\mathrm{C}} = \sum\nolimits_{i=1}^{N_{\mathrm{C}}} I_{\mathrm{C},i} \ , \tag{1}$$

where $I_{\mathrm{C},i}$ and N_{C} are the CCI received from i^{th} transmitter and the total number of the CCI, respectively.

Adjacent Channel Interference. When two LTE systems are operated in a neighboring spectrum, ACI cannot be avoided. There are two main sources of ACI: out-of-band emission and spurious emission [2]. ACI arises because of both transmitter imperfections and receiver imperfections [11]. High ACI will result in significant reduction in victim system capacity.

In a system simulation, the amount of interference is a key parameter. This can be expressed as Adjacent Channel Interference Ratio (ACIR), which is the measured total interference caused by a transmitter to an adjacent channel victim receiver. ACIR can be obtained as follows:

$$ACIR = \frac{1}{\frac{1}{ACLR} + \frac{1}{ACS}} \ . \tag{2}$$

For the transmitter, mostly as a result of transmitter non-linearities, the spectrum mask from the transmitter will leak into adjacent channels. It is characterized by the Adjacent Channel Leakage power Ratio (ACLR), which is defined as the ratio of the transmitted power to the power measured after a receiver filter in the adjacent RF channel [11]. For the receiver, it will have additional interference from the adjacent channel, since the receiver filter cannot be ideal [2]. It is characterized by the Adjacent Channel Selectivity (ACS), which is the ratio between the receiver filter attenuation on the assigned channel frequency and the receiver filter attenuation on the adjacent channel frequency [11].

From (2), it is evident that ACIR is dominated by the weakest ACLR/ACS. The minimum requirements for UE are 33 dB ACS and 30 dB ACLR for 10 MHz channel bandwidth [6]. The minimum requirements for BS are 43.5 dB ACS and 45 dB ACLR for 10 MHz channel bandwidth in a wide area [10]. Thus, the ACIR values for different links of 10MHz Channel bandwidth are given in Table 1:

Table 1. ACIR for Different Links

Link(10MHz)	**ACIR**(dB)
BS-UE	32.7
BS-BS	41.2
UE-BS	29.8
UE-UE	28.2

The ACI may be induced by the CCH or the data channel of the aggressor system. For simplicity, only the interference induced by the data channel in our ACI model is considered. It can be calculated in the following way:

$$I_{\mathrm{A}} = \sum\nolimits_{j=1}^{N_{\mathrm{A}}} I_{\mathrm{A},j} \quad , \tag{3}$$

where N_{A} is the total number of the ACI, and $I_{\mathrm{A},j}$ is the ACI received from the j^{th} transmitter which is reduced by ACIR dB.

Therefore, the Signal to Interference plus Noise Ratio (SINR) of a victim UE's CCH or data channel is expressed as:

$$SINR = \frac{S}{I_{\mathrm{C}} + I_{\mathrm{A}} + N} \quad . \tag{4}$$

where S is the desired signal power on the channel of the victim UE, and N is the thermal noise.

3 Mapping from CCHs Performance to Data Channels Throughput

In order to study how the data channels would be degraded when the CCHs are damaged by the ACI, we explored a simple model mapping from CCHs performance to data channels throughput. This model focuses on a mapping from the correct detection probability of CCHs to the throughput of data channels. In this way, the impact of ACI on CCHs is converted to system throughput loss.

3.1 Theoretical Analysis

The *Multiplication Axiom* in Probability Theory is expressed as:

$$P(A \cap B) = P(A|B)P(B) \quad , \tag{5}$$

where A and B are two events in the same probability space. $P(A \cap B)$ is the probability that A and B occur together, $P(A|B)$ is the probability of event A under the condition of occurrence of event B, and $P(B)$ is the probability that event B occurs. Using the definition of conditional probability, it follows from either formula that

$$P(A \cap B) = P(A)P(B) \ , \tag{6}$$

which is the definition of *statistical independence*. According to the *Multiplication Axiom* and *statistical independence* of incorrect bits in each channel, the ultimate throughput of data channels under consideration of incorrect detection probability of CCHs could be expressed as:

$$T' = P_{\text{related CCH1}} P_{\text{related CCH2}} \cdots P_{\text{related CCH}n} T \ , \tag{7}$$

where P is the correct detection probability of each related CCH ($P = 1 - \mathit{BLER}$), and T is the throughput of data channel, which is either PUSCH or PDSCH.

The quality of each data channel could be expressed by a mapping method from the SINR to the throughput. Such mapping could be approximated by an attenuated and truncated form of the Shannon bound, which is referred from 3GPP TR 36.942 [2]. For control channels, the link performance of CCHs is generally characterized by SINR-BLER curves, which decides whether the CCH information can be decoded correctly or not. The BLER of CCH is obtained via mapping table

$$\mathit{BLER} = f(\gamma_{\text{eff}}, \mathit{MCS}) \ , \tag{8}$$

where γ_{eff} is the effective SINR value of CCH and MCS is comprised by the selected modulation and coding rate. The link level SINR-BLER table is obtained from the link level simulation result.

3.2 Downlink and Uplink Throughput

The data channel only correlates with two control channels, i.e. PDCCH and PCFICH, in each subframe as each CCH has its own function in LTE system. For the other four control channels: PBCH is used to broadcast the basic system information when UE initially accesses the cell; PRACH carries the random access preamble, and has no effect on data channels; PHICH carries downlink ACK/NACK associated with uplink data transmission; and PUCCH supports combinations of Channel State Information (CSI) and HARQ. PHICH and PUCCH need not be considered since no HARQ method is adopted and the CSI is available for scheduling in coexistence simulation [2]. However, PDCCH carries scheduling assignments (resource allocation and MCS for both uplink and downlink) and other control information for a UE or group of UEs. PCFICH indicates the number of OFDM symbols used for control channel information in this subframe. Therefore, the information carried by the data channel can accurately transmit only under the condition that the information carried by PDCCH is error free. Furthermore, the information carried by PDCCH can be transmitted accurately only under the condition that the information carried by PCFICH is error free. In addition, PCFICH and PDCCH are tested jointly, i.e. a missing detection of PCFICH implies a missing detection of PDCCH [6]. Thus the ultimate throughput of PDSCH should be

$$T'_{\mathrm{PDSCH}} = P_{\mathrm{PDCCH/PCFICH}} T_{\mathrm{PDSCH}} \ , \tag{9}$$

and the ultimate throughput of PUSCH should be

$$T'_{\mathrm{PUSCH}} = P_{\mathrm{PDCCH/PCFICH}} T_{\mathrm{PUSCH}} \ . \tag{10}$$

4 Simulation Assumptions and Evaluation Methodology

The simulations in this study are based on the typical coexistence scenarios, as described in [2]. The TD-LTE and LTE-FDD systems are deployed in one layer of macro cellular network. The service area in simulation is a layout of 2-tier 19 hexagonal cells with tri-sector in each cell. Cell range is 500 meters, and the corresponding inter-site-distance (ISD) is 750 meters. Two systems are deployed in uncoordinated network, implying that the victim network's sites are located at the aggressor network's cell edges. Additionally, simulation frequency of two systems with 10 MHz bandwidth is 2.6 GHz, and the numbers of active UEs per sector are 5 for uplink and 50 for downlink. The other several important simulation assumptions can be found in reference [2], including propagation models, scheduling mechanism, and link level throughput mapping. In addition, power control scheme, ACLR model, and simulation methodology will be discussed in detail.

4.1 Power Control

As suggested by 3GPP RAN WG4, the fractional power control scheme [2] is adopted here. It is used to control the UE transmit power to compensate a fraction of the path coupling loss, as shown in the following equation:

$$P_t = P_{\max} \times \min\left\{1,\ \max\left[R_{\min},\ \left(\frac{CL}{CL_{x-\mathrm{ile}}}\right)^{\gamma}\right]\right\} \ , \tag{11}$$

where $P_{\max}$ is the maximum transmit power, $R_{\min}$ is the minimum power reduction ratio to prevent UEs with good channels to transmit at very low power level, CL is the path coupling loss (including antenna gain and shadow fading) for the UE and $CL_{x-\mathrm{ile}}$ is the x-percentile CL value. With this power control equation, the x percent of UEs that have the highest coupling loss will transmit at $P_{\max}$. Finally, $0 < \gamma \leq 1$ is the balancing factor for UEs with bad channel and UEs with good channel. In our simulation, two parameter sets for power control (PC) are used and specified in Table 2. In the downlink, fixed power per frequency resource block (RB) is assumed, and no power control mechanism is adopted [2].

Table 2. Power Control Algorithm Parameters

Parameter Set	γ	$PL_{x-\mathrm{ile}}$ (10 MHz)
Set 1	1	112-$\varDelta$
Set 2	0.8	129-$\varDelta$

Note:$\varDelta = 21\log_{10}(f_c/2.0)$, adjustment parameter related to different carrier frequency points. For $f_c = 2.6$ GHz, $\varDelta = 2.4$ dB.

4.2 ACLR Model

When we calculate the ACI, we should take the width and location of the resource unit within the frequency domain into account. For downlink, a common ACIR for all frequency resource blocks to calculate ACI should be used [2]. For uplink, the ACIR is dominated by the UE ACLR. According to [2], the ACLR model is described in Table 3. The ACLR model consists of two emission levels. If the aggressor UE is adjacent to the edge of victim RB, the frequency distance between the aggressor and the victim is not lager than 10RBs, and the ACLR value is assumed to be $30 + X$ dB. Otherwise, the ACLR value is assumed as $43 + X$ dB. X serves for simulations step with 5 dB step. For the CCHs, we consider the worst case scenario, whose allocated RBs are most near the aggressor system, and all affected RB of the CCHs will fall into the first level of ACLR model.

Table 3. ACLR Model for Uplink

Location of aggressor 10 RBs (10*180KHz)	**ACLR** (dB)
Adjacent to edge of victim (less than 10 RBs)	30+X
Non adjacent to edge of victim (more than 10 RBs away)	43+X

4.3 Simulation Methodology

Simulations which investigate the mutual interference impact of TD-LTE and LTE-FDD are based on snapshots. UEs are uniformly placed in the service area after deploying cell layout and setting simulation parameters, and then attached to servicing BS according to path loss including antenna gains and log-normal fading. Next, the intra-system and inter-system interference link's path loss including antenna gains and log-normal fading should be calculated. Then, the UEs will be scheduled according to their Channel Quality Indicator (CQI) based on resource scheduling strategy such as Round-Robin (RR) with full buffer traffic. After that, each UE will be allocated a certain amount of resource for their CCHs and data channel according to their C/I and QoS requirement. After resource allocation, power allocation strategy will start. The mechanism is presented in Sect. 4.1. Finally, the CCH SINR and data channel SINR of each UE will be calculated and collected, and the correct detection probability of CCH and throughput of the data channel per user can be obtained from the link level mapping models. Thus the ultimate throughput of the data channel can be calculated with (9) and (10).

5 Simulation Results and Discussions

Simulation results are presented in terms of throughput reduction percentage relative to the reference throughput without external interference, separately for all UE and for the 5% throughput cumulative distribution function (CDF) UE. From (9) and (10), the CCHs only consider the joint detection performance of the PDCCH and PCFICH. As suggested by 3GPP RAN WG4, three scenarios [6] are used to jointly verify the performance of PDCCH/PCFICH for different combinations of control channel elements (CCE) aggregation, antenna setup, channel bandwidth, and channel model. They are marked as 2CCE, 4CCE, and 8CCE, respectively.

In the downlink ACI scenario, Figs. 1 and 2 show the average downlink throughput loss and the 5% CDF downlink throughput loss, respectively. They indicate that the throughput loss with CCH is higher than that without CCH. When ACIR value is 32.7 dB, the biggest gap between the throughput loss with CCH and throughput loss without CCH is less than 0.5% for the average downlink throughput loss. For the 5% CDF downlink throughput loss, the biggest gap is about 4%, indicating that some edge UEs of victim may suffer low geometries of CCH coverage which lead to the unreliable reception of data, due to the strong ACI from the downlink of aggressor system to them. Average uplink throughput loss with CCH and without CCH are shown in Fig. 3, we find that the throughput loss in PC set 2 is higher than that in PC set 1, since the higher

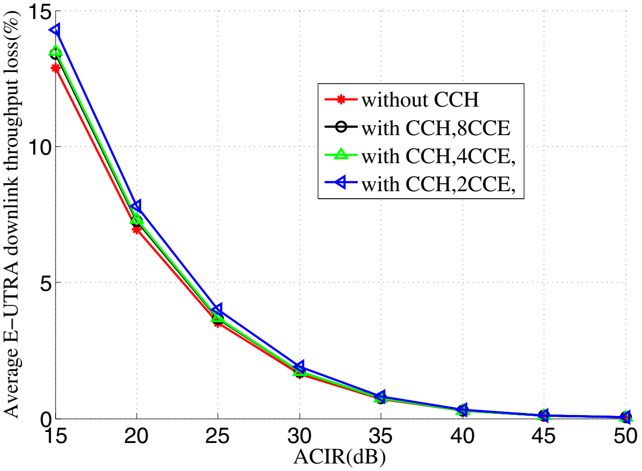


Fig. 1. Average E-UTRA downlink throughput loss, when ACI comes from downlink

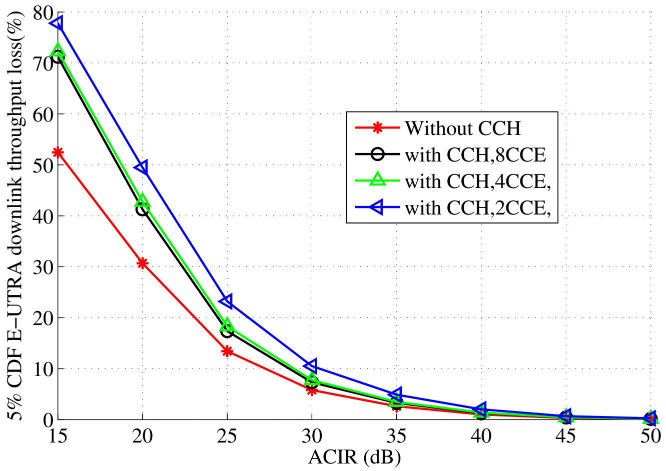


Fig. 2. 5% CDF E-UTRA downlink throughput loss, when ACI comes from downlink

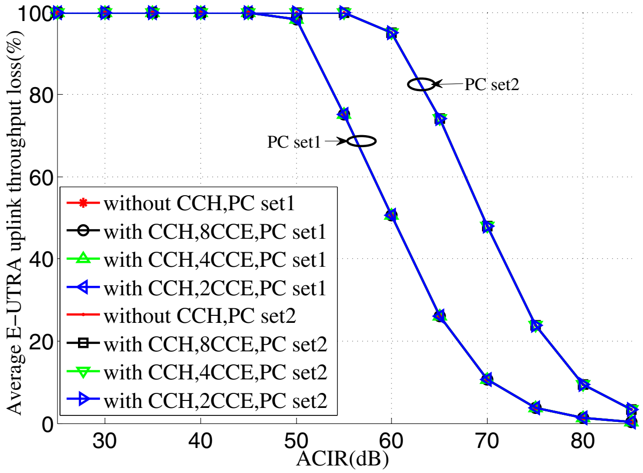


Fig. 3. Average E-UTRA uplink throughput loss, when ACI comes from downlink

transmitting power are used to overcome the interference in set 1. As shown in Fig. 3, the throughput loss without CCH and with CCH have overlapped results. The reason is: when ACIR ranges from 25dB to 50dB, uplink throughput for the coexistence of TD-LTE and LTE-FDD is equal to 0, so the PDCCH BLER does not have any impact on the throughput loss of PUSCH. When ACIR is bigger than 50 dB (i.e. the ACI is very weak), the PDCCH BLER in the coexistence scenario has little difference from that in the single system scenario. However, in this scenario, only when ACIR exceeds 85 dB, uplink throughput loss can drop to below 5%, that is, the damage degree of the downlink ACI impacting uplink throughput is more serious, the link between BSs need extra isolation.

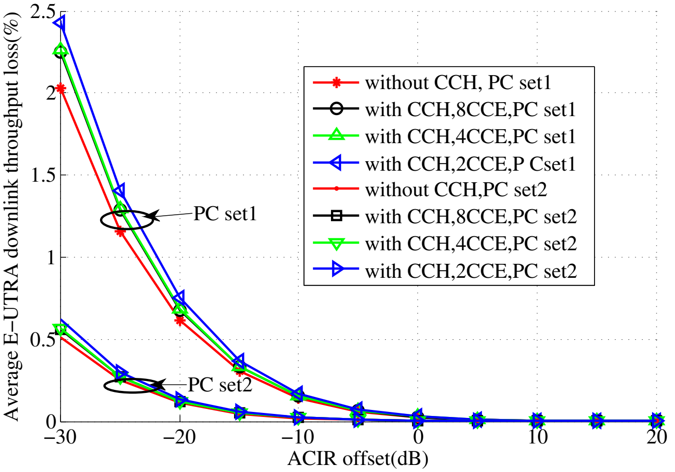


Fig. 4. Average E-UTRA downlink throughput loss, when ACI comes from uplink

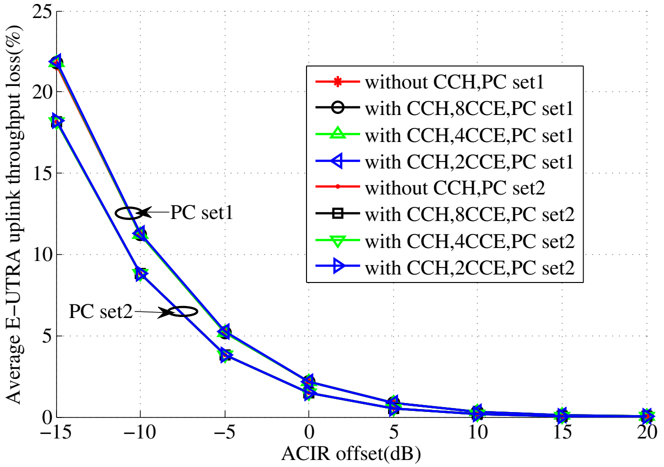


Fig. 5. Average E-UTRA uplink throughput loss, when ACI comes from uplink

In the uplink ACI scenario, Figs. 4 and 5 show the average downlink and uplink throughput loss, respectively. From fig. 4, we find that the average downlink throughput losses without or with CCH are very small. They are less than 2.5% for power control parameter set 1 and 2, when ACIR is 0 dB (ACIR offset is -30 dB), because the transmission power of aggressor UE is much less than that of the BS. Comparing different uplink throughput loss without or with CCH, Fig. 5 indicates that they overlap each other. The reason is: in this scenario, the uplink ACI is very weak, the interference on PDCCH/PHICH have little influence to the victim system performance. When the ACIR is 30 dB (i.e. the

Table 4. The Corresponding ACIR Value for 5% Throughput Loss

Interfering link	**Average /5% CDF**	**PC set**	**ACIR (dB)**		
			Without CCH	*With CCH,the largest*	*Extra required*
DL → DL	Average		22.83	23.67	0.84
	5% CDF		31.10	34.70	3.60
DL → UL	Average	1	74.0919	74.0921	0.0002
	Average	2	83.5883	83.5883	0*
	5% CDF	1	70.3205	70.3210	0.0005
	5% CDF	2	81.4041	81.4041	0*
UL → DL	Average	1	N/A**	N/A**	N/A
	Average	2	N/A**	N/A**	N/A
	5% CDF	1	1.86	7.56	5.70
	5% CDF	2	N/A**	N/A**	N/A
UL → UL	Average	1	25.31	25.39	0.08
	Average	2	23.94	23.98	0.04
	5% CDF	1	28.12	28.14	0.02
	5% CDF	2	26.27	26.31	0.04

Note*: The observed values are the same to four decimal places.
Note**: When the ACIR value is 0 dB, the throughput loss with/without CCH cannot meet the 5%.

ACIR value for link UE-BS), the uplink throughput loss is below 2.5% for power control parameter set 1/2.

In general, the evaluation criterion of system performance on the coexistence scenario is that the throughput loss of the victim system shall not exceed 5%. Thus, the corresponding ACIR values for 5% throughput loss in four interference scenarios, which are very high when downlink interferes with uplink scenario, are shown in Table 4. For the uplink ACI scenarios, the throughput loss without/with control channels is very small because the uplink ACI is very weak. Therefore, the degree of isolation between BSs needs to be upgraded from 41.2dB to 83.6dB.

6 Conclusion

Throughput performance evaluation with/without CCH is presented under different coexistence scenarios. A mapping from CCHs performance to data channels throughput and simulation results have been discussed. The work of this paper will be useful for LTE deployment.

Acknowledgments. This work is supported by DOCOMO Beijing Communications Laboratories Co., Ltd., National Basic Research Program of China (2012CB316000), NSFC (61021001), National S&T Major Project (2011ZX03005-003-01) and Tsinghua Univ. Initiative Scientific Research Program (2010THZ03-02).

References

1. Global mobile Suppliers Association: Evolution to LTE Report (2013), `http://www.gsacom.com/`
2. 3GPP TR 36.942: Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios. Technical report, V10.3.0 (2012)
3. Ruiming, Z., Xin, Z., Xi, L., Qun, P., Yinglong, F., Dacheng, Y.: Performance evaluation on the coexistence scenario of two 3gpp lte systems. In: Vehicular Technology Conference (VTC 2009-Fall), pp. 1–6. IEEE (2009)
4. Chen, X., Jin, X., Moorut, P., Love, R., Sun, Y., Xiao, W., Fernandes, E.: Coexistence Analysis Involving 3GPP Long Term Evolution. In: Vehicular Technology Conference (VTC-2007 Fall), pp. 225–229. IEEE (2007)
5. Jia, H., Miao, Q., Yin, C., Wu, H., Ma, Y.: Performance analysis of coexistence between LTE-TDD and TD-SCDMA. In: Communications Technology and Applications (ICCTA 2009), pp. 303–307. IEEE (2009)
6. 3GPP TS 36.101: Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception. Technical report, V10.4.0 (2011)
7. 3GPP TS 36.211: Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation. Technical report, V10.3.0 (2011)
8. 3GPP TS 36.213: Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures. Technical report, V10.3.0 (2011)
9. Sesia, S., Toufik, I., Baker, M.: LTE: the UMTS long term evolution. John Wiley & Sons, New York (2009)
10. 3GPP TS 36.104: Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception. Technical report, V10.4.0 (2011)
11. 3GPP TR 25.942: Technical Specification Group Radio Access Networks; Radio Frequency (RF) system scenarios. Technical report, V10.0.0 (2011)

Buffer Occupation in Wireless Social Networks

Tuo Yu, Xiaohua Tian, Feng Yang, and Xinbing Wang

Department of Electronic Engineering,
Shanghai Jiao Tong University, Shanghai, China
{yutuo,xtian,yangfeng,xwang8}@sjtu.edu.cn

Abstract. In this paper, we investigate the buffer occupation in wireless social networks. n source nodes are randomly distributed in the networks, and each of them chooses several friend nodes, whose number follows power-law distribution. Two different strategies are proposed to achieve optimal buffer occupation or throughput, and the upper bound and lower bound for buffer size and throughput are also investigated. The results show that there exists trade-off between buffer occupation and throughput, and we also find that the buffer occupation for each node will not increase when the size of network becomes larger, which is opposite to our intuitive understanding.

1 Introduction

Online social networks (OSNs) have received great attention in the past several years. Tens of millions of users have been attracted by MySpace, Facebook, Orku, LiveJournal, Cyworld and Flickr. As a special kind of online social networks, wireless social networks are rapidly drawing research interests. For example, in [1] Li *et al.* investigated the capacity of ad hoc wireless networks where the distance between source-destination peers follows specific social characterization. In [2] Azimdoost *et al.* showed us the capacity of the networks where nodes communicate only with their long-range social contacts. Moreover, there are also a large number of works focusing on mobile social service [3–5], and geosocial networking [6, 7].

However, most of previous works assume that the buffer of node is infinite, and the buffer occupation in wireless social networks has not been investigated so far. Since the unstable connection in wireless networks, buffer size is necessary. Packets have to be stored in the relay node if the connection to the next neighbor is not available. Since the resource such as storage space is always limited in wireless networks [8], the study about buffer occupation is crucial in large scale social networks. In [9], J. D. Herdtner and E. Chong investigated the throughput capacity of mobile wireless networks with limited node buffer. In [10], S. Bodas *et al.* have studied scheduling methods in multichannel wireless networks in the small-buffer regime. In [11], Wang *et al.* has studied the achievable lower bound for buffer size occupied in intermittently connected static wireless networks.

In this paper, we are motivated to present a first look into the buffer occupation in wireless social networks from a theoretical perspective. Different from the case in normal wireless networks, there exist specific types of interdependency between users. For instance, the statistical information from numerous social networks such as Youtube

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 25–41, 2013.

shows an interesting feature that the distribution of nodes follows power-law degree distribution [12, 13]. In this paper, we will investigate the effect of this feature on the buffer of nodes, and the results can also be extended to other networks with social characteristics [14–17].

Moreover, we also investigate the trade-offs between buffer occupation and throughput. In [18], the authors analyzed the buffer size basing on the assumption that the transmission rate of source node is constant, and the bandwidth between different peers is infinite, which is not practical in real wireless networks. [19] investigated trade-offs in buffer requirements and throughput constraints for synchronous dataflow graphs, however the result is not applicable in wireless social networks. In wireless social networks, the performance of throughput is also different from that in typical wireless networks, which leads to another challenge in our work. To investigate the effect of throughput on buffer size, we propose two different transmission strategies (π_1 and π_2), and highways system based on percolation theory is also used to improve the performance of throughput.

This paper is organized as follows. In Section 2, we present the network model and some basic assumptions. In Section 3, we propose the first transmission strategy π_1 and analyze the buffer occupation and throughput achieved by this strategy. In Section 4, we propose the second strategy π_2 which is based on percolation theory, and its performance is also analyzed. In Section 5, we investigate the lower bound of buffer occupation and in Section 6 the upper bound of throughput is analyzed. We discuss the results in Section 7 and draw our conclusion in Section 8.

2 Wireless Network Model

We mainly focus on the random extended network (REN). The network covers a square with size $[0, \sqrt{n}] \times [0, \sqrt{n}]$, and nodes are placed according to a Poisson point process (p.p.p.) of unit intensity. According to Chebyshevs Inequality, it can be easily obtained that the number of nodes is within $((1 - \varepsilon_0)n, (1 + \varepsilon_0)n)$, where ε_0 is a positive infinitesimal. Therefore we can simply assume that the number of nodes is n, which will not impact our results in order sense.

2.1 Channel Model

Let d_{ij} denote the Euclidean distance between node i and j. When node i is transmitting to node j, the reception power is denoted by P_{ij}. We assume that P_{ij} follows the power propagation model

$$P_{ij} = C\frac{P_i}{d_{ij}^{\gamma}}, \tag{1}$$

where P_i is the transmission power, γ is the path loss exponent, and C is a constant related to the antenna profiles of the transmitter and the receiver, wavelength, and so on. We assume that $\gamma > 2$, which has been widely used in the study of outdoor wireless transmission. For the transmission rate between node i and j, we apply Gaussian channel model, which is denoted by propagation model

$$R_{ij} = W \log(1 + SINR_{ij}), \tag{2}$$

where W is the maximum channel bandwidth, and $SINR_{ij}$ is the Signal-to-Interference plus Noise Ratio between node i and j, which is calculated by

$$SINR_{ij} = \frac{P_{ij}}{N + \sum_{k \neq i} P_{kj}}, \tag{3}$$

in which N is the background noise received at j, and $\sum_{k \neq i} P_{kj}$ is the interference from all the other transmitting nodes. In this paper we assume that all the nodes have the same transmission power, which is denoted by P.

2.2 Social Relations and Traffic Pattern

The n nodes in the system form a social network. For a node $i \in [1, n]$, it chooses K_i friend nodes randomly from the rest nodes in the network, and these $K_i + 1$ nodes form a multicast group G_i. K_i is a random variable which follows power-law degree distribution [12, 13]. That is

$$Pr\{K_i = k\} = \frac{1}{Ak^{\alpha}},$$

where α is the power-law parameters, $A = \sum_{j=1}^{n} \frac{1}{j^{\alpha}}$ is the normalizing factor.

The major traffic pattern we study is social multicast, i.e. information is broadcasted to all friends of the source. We assume that all the multicast group do not change with time, and the packet arrive rate at each source is r_m. In our network packets are transmitted by multi-hop, and if a packet arriving a node cannot be transmitted immediately, it will be buffered at the node awaiting the next transmission opportunity.

2.3 Definition of Throughput

Feasible social multicast throughput: Per-node throughput $\Lambda(n)$ is said to be feasible if there is a spatial and temporal scheme for scheduling transmissions, such that by operating the network in a multi-hop fashion and buffering at intermediate nodes when awaiting transmission opportunities, every source can send $\Lambda(n)$ bits/sec to its K_i chosen destination nodes. That is, there is a $T < \infty$ such that in every time interval $[(i-1)T, iT]$, every source can sent $T\Lambda(n)$ bits to each of its K_i destinations.

2.4 Definition of Buffer Occupation

Before we define buffer occupation, it is necessary to make the following assumption about the processing speed of nodes.

Assumption on Processing Speed: We assume that time in the network is slotted with a constant slot length T_s, which is a constant. Compared with the length of time slot T_s, the processing speed of every node is short enough to be ignored. This assumption helps us focus on the effects of different transmission strategies on the throughput and buffer occupation, and other factors are considered ideal.

Buffering: Any message received by a relay node during a time slot will be buffered before they are transmitted to the next relay node or the destination at the next time slot. Therefore the size of the messages buffered in a node is related to the length of a time slot T_s.

Buffer Occupation: According to the definition of buffering, the buffer occupied at each node is maximal at the end of each time slot. We define buffer occupation $S(n)$ as the average buffer occupation of one node at the end of each time slot in the network. Since nodes are placed according to a Poisson point process (p.p.p.) of unit intensity, it is reasonable to consider the average value of buffer occupation [11].

It is obvious that $S(n)$ is related to r_m, e.g. $S(n) = 0$ when $r_m = 0$. Therefore to focus on the buffer occupation, we temporarily assume that r_m is a constant, and let $S_r(n)$ denote the buffer occupation without regard to r_m. We will discuss the effect of r_m on $S(n)$ when $r_m = \Lambda(n)$ in a latter section and thus investigate the relationship between throughput and buffer occupation.

Packets: We call the messages generated by a source during one time slot a packet. According to the assumption above, the size of one packet is $r_m T_s$.

Data Flow: The transmission paths originating from the same source node and the packets transmitted on them are called as the data flow from the source node.

3 Buffer Occupation and Throughput Achieved by Strategy π_1

In this section we propose a transmission strategy (denoted by π_1) that minimizes the buffer occupation in our network. Meanwhile we will also analyze the throughput achieved by this strategy. The result of this section is shown by Theorem 1.

Theorem 1. *The buffer occupation achieved by π_1 satisfies*

$$S_r^{(1)}(n) = \begin{cases} O(r_m T_s \sqrt{n}/\sqrt{\log n}) & , \quad \alpha > 3/2 \\ O(r_m T_s \sqrt{n \log n}) & , \quad \alpha = 3/2 \\ O(r_m T_s n^{2-\alpha}/\sqrt{\log n}) & , 1 < \alpha < 3/2 \\ O(r_m T_s n/(\log n)^{3/2}) & , \quad \alpha = 1 \\ O(r_m T_s n/\sqrt{\log n}) & , \quad 0 \leq \alpha < 1 \end{cases} \tag{4}$$

The throughput achieved is

$$\Lambda^{(1)}(n) = \begin{cases} \Omega(1/\sqrt{n \log n}) & , \quad \alpha > 3/2 \\ \Omega(1/\log^{3/2} n\sqrt{n}) & , \quad \alpha = 3/2 \\ \Omega(n^{\alpha-2}/\sqrt{\log n}) & , 1 < \alpha < 3/2 \\ \Omega(\sqrt{\log n}/n) & , \quad \alpha = 1 \\ \Omega(1/n\sqrt{\log n}) & , \quad 0 \leq \alpha < 1 \end{cases} \tag{5}$$

We will prove Theorem 1 in the following subsections.

3.1 Transmission Strategy π_1

The main idea of Strategy π_1 is that, for each multicast group G_i, we firstly construct a spanning tree using Prims algorithm, and then choose transmission path for each edge to construct an optimal routing tree $ORT(G_i)$. Note that even though the number of nodes in G_i is $K_i + 1$, we recognize it as K_i to simplify our analysis.

Strategy π_1 for $ORT(G_i)$:

(a)Construction of spanning tree:

(1)The K_i nodes of G_i belong to K_i different components initially. Set $z = 1$.

(2)Divide the network into at most $K_i - z$ different squares, and the edge of each square is $\sqrt{n}/\lfloor\sqrt{K_i - z}\rfloor$.

(3)If a square contains at least two nodes from different components, then link the nearest two nodes that come from two different components. Thus these components are combined to be a new component.

(4)If $z = K_i - 1$, return this spanning tree. Otherwise set $z = z + 1$ and goto step(2).

One example of constructing a spanning tree is shown by Fig.1.

(b)Choice of transmission path:

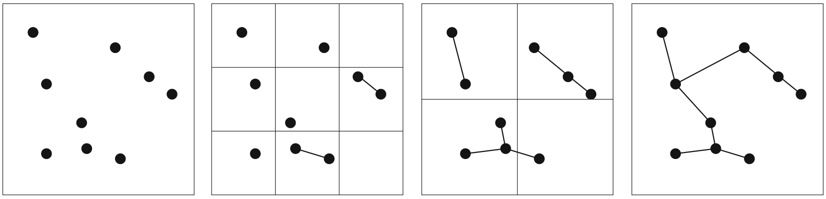
Divided the network into grids with edge length Ψ. For each edge in the spanning tree $v_i \to v_j$, implement the following steps.

(1)Assume that v_i is located in grid s_i and v_j is in grid s_j. The packets from v_i will be firstly transmitted along the squares that have the same x-coordinate as s_i until it reaches a square that have the same y-coordinate as s_j. For each passed grid, randomly choose a node from the grid to act as a relay node.

(2)The packets are transmitted along the squares that have the same y-coordinate as s_j until it reaches s_j. Finally the packets are transmitted to v_j.

(c)Remove cycles if any. Then return the optimal routing tree $ORT(G_i)$.

For all the multicast group in the network, implement Strategy π_1. To ensure the connectivity of the nodes, we set $\Psi = \Theta(\log n)$, which is confirmed by Lemma 1.



(a) Origin (b) $z = 1$ (c) $z = 2$ (d) $z = 9$

Fig. 1. Example for the construction of spanning tree

Lemma 1. *Each grid contains at least one node with high probability (w.h.p.) when* $\Psi = \Omega(\sqrt{\log n})$.

Proof. Divide the network into grids with edge $l = \sqrt{b_1 \log n}$, where b_1 is a constant and $b_1 > 1$. For grid s_i, the probability that it contain at least one node is

$$P_i = 1 - (1 - \frac{l^2}{(\sqrt{n})^2})^n = 1 - e^{n \log 1 - \frac{b_1 \log n}{n}} = 1 - \frac{O(1)}{n^{b_1}} \to 1,\ n \to \infty.$$

Let n_s denote the number of grids, then $n_s = (\sqrt{n})^2/l^2 = n/b_1 \log n$. Therefore the probability that each grid contains at least one node is

$$P_n = P_i^{n_s} = (1 - \frac{O(1)}{n^{b_1}})^{n_s} = e^{-\frac{O(1)}{b_1 n^{b_1 - 1} \log n}}.$$

Since $b_1 > 1$, $P_n \to \infty$ when $n \to \infty$. Thus with the condition that $\Psi = \Omega(\sqrt{\log n})$, each grid contains at least one node w.h.p. when the number of nodes in the network is large enough.

Lemma 1 confirms the connectivity of our network, and the following lemma will show that the sum transmission rate of all the paths going through a grid can achieve a constant order.

Lemma 2. *The sum transmission rate of all the paths going through a single grid is at least b_2W, where b_2 is a confirmable positive constant.*

Proof. We apply a transmission method similar to 9-TDMA. Every nine neighbouring grids constitute a grid group. Label the grids in each group with number 1 to 9. Divide each time slot to nine sub-slot, and label them with $t = 0, 1, 2, ..., \infty$. At sub-slot t, all the grids with label $(t \ mod \ 9) + 1$ are allowed to transmit packets simultaneously. Otherwise they buffer packets in their relay nodes. Therefore when grid s_i is active, the other grids interfering s_i uniformly distribute on the squares centered at s_i. For the jth square there are $8j$ interfering grids, and the distance to s_i is at least $(3j - 2)\Psi$. Thus the interference received by s_i is

$$I_i \leq \sum_{j=1}^{\infty} 8j \times \frac{CP}{[(3j-2)\Psi]^\gamma} \leq \frac{8CP}{\Psi^\gamma}[1 + \sum_{j=2}^{\infty}(3j-2)^{1-\gamma}] < \frac{8CP(3\gamma-5)}{\Psi^\gamma(3\gamma-6)}.$$

Denote the signal power received at s_i from one of its neighbouring grids by D_i. Since the maximum distance between a transmitter and a receiver is no larger than $\sqrt{5}\Psi$, D_i satisfies $D_i \geq CP/(\sqrt{5}\Psi)^\gamma$. Then for the Signal-to-Interference plus Noise Ratio at the receiver we have

$$SINR_i = \frac{D_i}{N + I_i} \geq \frac{CP/(\sqrt{5}\Psi)^\gamma}{N + 8CP(3\gamma-5)/\Psi^\gamma(3\gamma-6)}.$$

Let $P = c_1\Psi^\gamma$, where c_1 is a positive constant. Therefore we have

$$SINR_i \geq \frac{c_1 C}{5^{\frac{\gamma}{2}}(N + 8c_1C(3\gamma-5)/(3\gamma-6))}.$$

Thus it is clear that $SINR_i$ has a lower bound that is independent of n. According to the definition of R_{ij}, it is confirmed that for every time slot, any grid can transmit packets with at least a constant data rate. We let b_2W denote the transmission rate that any grid can achieve, where b_2 is a finite positive constant.

3.2 Buffer Occupation Achieved by Strategy π_1

Before we investigate the buffer occupation achieved by π_1, it is necessary to introduce the following famous lemma.

Lemma 3. *[20] Let node Y_i, $1 \leq i < \infty$ be independent and identically distributed random variables in $\mathbb{R}^d$, $d \geq 2$. Denote $M_k = |EMST(\{Y_1, ..., Y_k\})|$ as the length of the Euclidean minimum spanning tree which consists of $Y_1, ..., Y_k$. Then with probability* 1*,*

$$\lim_{k\to\infty} M_k = c(d)n^{(d-1)/d}\int_{\mathbb{R}^d} f(x)^{(d-1)/d}dx,$$

where $f(x)$ is distribution density function of Y_i, and $c(d)$ is the constant that is independent of k.

For this paper we set $f(x) = 1/n$ and $d = 2$, and then we have $M_k = \Theta(\sqrt{kn})$. Since for each multicast group K_i follows power-law degree distribution, we can calculate the expected length of the Euclidean minimum spanning tree for a multicast group.

$$E\{\mid EMST(G_i) \mid\} \sim \frac{1}{A}\sum_{k=1}^{n-1} \frac{\sqrt{kn}}{k^{\alpha}} \sim \begin{cases} \sqrt{n} & , \quad \alpha > 3/2 \\ \log n\sqrt{n} \ , & \quad \alpha = 3/2 \\ n^{2-\alpha} & , 1 < \alpha < 3/2 \\ n/\log n \ , & \quad \alpha = 1 \\ n & , \quad 0 \leq \alpha < 1 \end{cases} . \tag{6}$$

The following lemma shows us the number of multicast trees that go through a single grid.

Lemma 4. *For a grid s_j, the upper bound for the probability that it is passed by $ORT(G_i)$ is $c_2\sqrt{\Psi}E\{|EMST(G_i)|\}/n$, where c_2 is constant.*

Proof. During the construction of $ORT(G_i)$, let Boolean variable F_z denote whether $ORT(G_i)$ passes s_j at the zth step, i.e. when s_j is passed by $ORT(G_i)$ at the zth step, $F_z = 1$, otherwise $F_z = 0$. Through geometrical deduction we have

$$Pr\{F_z = 1|K_i\} = p_i(z)\frac{n/(K_i - z)}{n} = p_i(z)\frac{1}{K_i - z},$$

where $\frac{n/(K_i-z)}{n}$ is the probability that with edge length $\sqrt{n}/\sqrt{K_i - z}$, the square containing s_j is selected at the zth step. $p_i(z)$ denotes the probability that s_j is selected during the construction of the spanning tree. Assume that s_j lies in the pth row and the qth rank among the grids covered by the square, then we have

$$\begin{aligned} p_i(z) = {} & (p-1)\Psi^{\frac{3}{2}}(\frac{K_i - z}{n})^{\frac{3}{2}}(\frac{\sqrt{n}}{\sqrt{\Psi(K_i - z)}} - p + 1) \\ & +(q-1)\Psi^{\frac{3}{2}}(\frac{K_i - z}{n})^{\frac{3}{2}}(\frac{\sqrt{n}}{\sqrt{\Psi(K_i - z)}} - q + 1) \\ & \leq 2\sqrt{\Psi(K_i - z)}/\sqrt{n}. \end{aligned}$$

Thus the probability that s_j is passed by $ORT(G_i)$ is

$$\begin{aligned} Pr\{ORT(G_i) \text{ passes } s_j\} & \leq \sum_{z=1}^{K_i-1} Pr\{F_z = 1|K_i\} \leq \sum_{k=1}^{n-1}\sum_{z=1}^{k-1} \frac{2\sqrt{\Psi(K_i - z)}}{(K_i - z)\sqrt{n}} Pr\{K_i = k\} \\ & \leq \frac{4\sqrt{2\Psi}}{n}\sum_{k=1}^{n-1}\sqrt{nk}Pr\{K_i = k\} = c_2\sqrt{\Psi}E\{|EMST(G_i)|\}/n. \end{aligned}$$

Lemma 4 shows the probability for a give grid to be passed by a given multicast tree. Further more, we analyze the number of multicast trees that go through a single grid.

Lemma 5. *Denote $M(s)$ as the number of multicast trees that go through a grid s, and*

$$\Pi \triangleq \{M(s) \leq 2c_2\sqrt{\Psi}E\{|EMST(G_i)|\} \text{ holds for any grid}\}. \tag{7}$$

Then Π will happen with possibility 1 when $n \to \infty$.

Proof. Let Boolean variable F_{si} denote whether the grid s is passed by $ORT(G_i)$. When it is true, $F_{si} = 1$, otherwise $F_{si} = 0$. For different grid, F_{si} is an i.i.d. Bernoullian random variable. According to the definition of $M(s)$, we have $M(s_j) = \sum_{i=1}^{n} F_{si}$. Denote the excepted value of F_{si} by p_1. According to Lemma 4, we have $p_1 \leq c_2\sqrt{\Psi}E\{|EMST(M_i)|\}/n \triangleq p_2$. Let $M^*(s)$ denote the sum of n i.i.d. Bernoullian random variables with mean value p_2. Thus $M^*(s) \geq M(s)$ with high probability. According to Chernoff bounds we have

$$Pr\{M(s) > 2p_2\} \leq Pr\{M^*(s) > 2p_2\} < (e/4)^{np_2} < e^{-np_2/8}.$$

Since we have gotten $E\{|EMST(M_i)|\} \geq \Theta(\sqrt{n})$ and $\Psi = \Omega(\log n)$,

$$Pr\{\Pi \text{ is true}\} \geq 1 - \sum_{s} Pr\{M(s) > 2p_2\} > 1 - e^{-\sqrt{n \log n}/8} \to 1,\ n \to \infty.$$

Since the transmission between two neighbouring grids follows the method similar to 9-TDMA, the delay for a single hop does not exceed $2T_s$. Thus for one grid the buffer occupied by a single mulitcast tree will not exceed $2r_mT$. According to Lemma 4, the buffer occupied in one grid is no larger than $2r_mT_sM(s)$. Based on the fact that the number of grids in the network is $n/\log n$, and the number of nodes is n, we get the buffer occupation for a single node

$$S_r^{(1)} = 2r_mT_sM(s) \cdot \frac{n}{n \log n} < \frac{4r_mT_sc_2}{\sqrt{\log n}} E\{|EMST(M_i)|\}. \tag{8}$$

Combining (6) and (8), we can get (4) in Theorem 1 directly.

3.3 Throughput Achieved by Strategy π_1

According to Lemma 2, the data rate between two neighbouring grids can achieve constant order, which is shared by all the data flows that go through this path. By Lemma 5, the throughput of the network is $\Omega(1/\sqrt{\Psi}E\{|EMST(M_i)|\})$, which leads to (5) in Theorem 1 directly.

4 Buffer Occupation and Throughput Achieved by Strategy π_2

In this section we propose another transmission strategy (denoted by π_2) that maximize the throughput in our network. The buffer occupation achieved by this strategy is also analyzed. The mainly result of this section is shown by Theorem 2.

Theorem 2. *The buffer occupation achieved by* π_2 *satisfies*

$$S_r^{(2)}(n) = \begin{cases} O(r_mT_s\sqrt{n}) & ,\quad \alpha \geq 2 \\ O(r_mT_s(\sqrt{n} + n^{2-\alpha}\log n)) & ,\ 3/2 < \alpha < 2 \\ O(r_mT_s\sqrt{n}\log n) & ,\quad \alpha = 3/2 \\ O(r_mT_sn^{2-\alpha}\log n) & ,\ 1 < \alpha < 3/2 \\ O(r_mT_sn) & ,\quad \alpha = 1 \\ O(r_mT_sn\log n) & ,\ 0 \leq \alpha < 1 \end{cases} \tag{9}$$

The throughput achieved is

$$\Lambda^{(2)}(n) = \begin{cases} \Omega(1/\sqrt{n}) & ,\quad \alpha \geq 2 \\ \Omega(1/(\sqrt{n} + n^{2-\alpha}\log n)) & ,\ 3/2 < \alpha < 2 \\ \Omega(1/\sqrt{n}\log n) & ,\quad \alpha = 3/2 \\ \Omega(1/n^{2-\alpha}\log n) & ,\ 1 < \alpha < 3/2 \\ \Omega(1/n) & ,\quad \alpha = 1 \\ \Omega(1/n\log n) & ,\ 0 \leq \alpha < 1 \end{cases} \tag{10}$$

We will prove Theorem 2 in the following subsections.

4.1 Transmission Strategy π_2

Before we propose π_2, it is necessary to introduce the preliminaries of *highway* [21]. The construction of highway in wireless transmission systems is based on percolation theory [22, 23]. The method we use to construct highway is similar to that in [21]. Firstly we slantly partition the network into n/c^2 subsquares with constant side length c. Thus there are $m = \lceil \sqrt{n}/\sqrt{2}c \rceil$ subsquares intersecting with a edge of the network. If a subsquare contains more than one node, we call it *open subsquare*, and its diagonal lines are called *open edges*. If a path consisting of open edges cross the entire network area vertically or horizontally, it is called an *open path*. If a subsquare is crossed by an open path, we randomly select a node from the subsquare to act as a relay node. Percolation theory confirms that the number of relay nodes in the network is the same order of $\Theta(n)$. All the open paths in the network construct the highways system.

If we partition the network into vertical or horizontal slabs of size $\sqrt{n} \times (\kappa \log m - \varepsilon_m)$, where κ is constant and ε_m ensures that the number of slabs is an integer, percolation theory confirms the following lemma.

Lemma 6. *[21] Let N^h (or N^v) denote the number of independent open paths in a horizontal (or vertical) slabs respectively. The independent open paths have no common relay nodes and are covered by the same slab. For any κ and p satisfying $2 + \kappa \log 6(1-p) < 0$, $p \in (5/6, 1)$, there exists a constant $\delta(\kappa, p)$ confirming that*

$$\lim_{m\to\infty} Pr\{N^h \geq \delta \log m\} = 1, \quad \lim_{m\to\infty} Pr\{N^v \geq \delta \log m\} = 1.$$

Note that $\delta(\kappa, p)$ is independent of n.

According to Lemma 6, there exist at least $\delta \log m$ independent open paths in a single slab w.h.p.. We further divide each slab into $\delta \log m$ slices of size $\sqrt{n} \times (\kappa \log m - \varepsilon_m)/\delta \log m$. Thus every slice has its own corresponding independent open path, and each node in the slice also corresponds to this path. Moreover, the distance between the node and its corresponding path is no larger than $\kappa \log m - \varepsilon_m$. Note that each node has both the horizontal and vertical corresponding paths.

Now we can propose the transmission strategy π_2. The main idea of π_2 is that packets are transmitted along the highways system to achieve a higher throughput. To distinguish the optimal routing tree based on highways system from $ORT(G_i)$, we let $ORTH(G_i)$ denote the optimal routing tree constructed by π_2. *Strategy π_2 for $ORTH(G_i)$:*

(a)Construction of spanning tree:

This part is similar to that in π_1-(a), and thus we omit it here.

(b)Choice of transmission path:

For each edge in the spanning tree $v_i \to v_j$, implement the following steps.

(1)The packets from v_i are firstly transmitted to the nearest node on its corresponding horizontal highway by one hop.

(2)The packets are transmitted along the horizontal highway until it reaches the subsquare where the corresponding vertical highway of v_j go through. Then the packets are transmitted to the nearest node on the vertical highway.

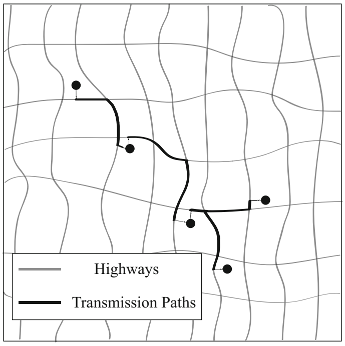


Fig. 2. Example for the construction of optimal routing tree

(3)The packets go long the vertical highway until it reach the node that is nearest to v_j. Then the node transmits the packets to v_j by one hop.

(c)Remove cycles if any. Then return the optimal routing tree $ORTH(G_i)$. One example is shown in Fig.2.

For all the multicast group in the network, implement Strategy π_2. Similar to that in π_1, after partitioning the network into grids, the transmission rate between two adjacent grids can achieve constant order.

4.2 Buffer Occupation Achieved by Strategy π_2

We firstly analyze the probability that a single subsquare is crossed by a give multicast tree. The result is given by Lemma 7.

Lemma 7. *For a given subsquare s_j^*, the probability for it to be crossed by $ORTH(G_i)$ is*

$$Pr\{ORTH(G_i)\,passes\,s_j^*\} = \begin{cases} O(1/\sqrt{n}) & , \quad \alpha \geq 2 \\ O(1/(\sqrt{n} + n^{1-\alpha}\log n)) & , 3/2 < \alpha < 2 \\ O(\log n/\sqrt{n}) & , \quad \alpha = 3/2 \\ O(n^{1-\alpha}\log n) & , 1 < \alpha < 3/2 \\ O(1) & , \quad \alpha = 1 \\ O(\log n) & , \quad 0 \leq \alpha < 1 \end{cases} \tag{11}$$

Proof. During the construction of $ORTH(G_i)$, let Boolean variable F_z^* denote whether $ORTH(G_i)$ passes s_j^* at the zth step, i.e. when s_j^* is passed by $ORTH(G_i)$ at the zth step, $F_z^* = 1$, otherwise $F_z^* = 0$. Similar to that during the proof for Lemma 4, we have

$$Pr\{F_z^* = 1|K_i\} = \frac{n/(K_i - z)}{n} \cdot p_i(z) = \frac{1}{K_i - z} \cdot p_i(z), \tag{12}$$

where $\frac{n/(K_i-z)}{n}$ is the probability for the square containing s_j^* to be selected at the zth step, and $p_i(z)$ denotes the probability for s_j^* to be crossed by transmission route. For $p_i(z)$, as shown in Fig.3, we have

$$p_i(z) \leq \frac{2c[\sqrt{\frac{n}{K_i-z}} + 2(\kappa\log m + \sqrt{2}c)]}{[\sqrt{\frac{n}{K_i-z}} + 2(\kappa\log m + \sqrt{2}c)]^2} = 2c/[\sqrt{\frac{n}{K_i - z}} + 2(\kappa\log m + \sqrt{2}c)]. \tag{13}$$

Combining (12) and (13), we have

$$Pr\{F_z^* = 1|K_i\} \leq \frac{1}{K_i - z} \cdot \frac{2c}{[\sqrt{\frac{n}{K_i - z}} + 2(\kappa \log m + \sqrt{2}c)]} \leq \frac{2c}{n}[\sqrt{\frac{n}{K_i - z}} + 2(\kappa \log m + \sqrt{2}c)].$$

The the probability that s_j^* is crossed by $ORTH(G_i)$ is

$$\begin{aligned} & Pr\{ORTH(G_i) \text{ passes } s_j^*\} \\ \leq & \sum_{z=1}^{K_i - 1} Pr\{F_z^* = 1|K_i\} \\ \leq & \sum_{k=1}^{n-1}\sum_{z=1}^{k-1} \frac{2c}{n}[\sqrt{\frac{n}{K_i - z}} + 2(\kappa \log m + \sqrt{2}c)]Pr\{K_i = k\} \\ \leq & \sum_{k=1}^{n-1}[\frac{4c}{n}\sqrt{2kn} + \frac{k-1}{n}(2c\kappa \log m + 2\sqrt{2}c^2)]Pr\{K_i = k\} \\ = & \frac{4\sqrt{2}c}{n}E\{EMST(G_i)\} + \frac{1}{n}(2c\kappa \log m + 2\sqrt{2}c^2)\sum_{k=1}^{n-1}(k-1)Pr\{K_i = k\}. \end{aligned} \tag{14}$$

Since we have calculated the order of $E\{|EMST(G_i)|\}$ in (6), and

$$\sum_{k=1}^{n-1}(k-1)Pr\{K_i = k\} \sim \begin{cases} 1 & , \quad \alpha \geq 2 \\ \log n & , \quad \alpha = 2 \\ n^{2-\alpha} & , 1 < \alpha < 2 \\ n/\log n & , \quad \alpha = 1 \\ n & , 0 \leq \alpha < 1 \end{cases} \tag{15}$$

by combining (6), (14) and (15) we can derive (11).

Denote the number of multicast trees that go through subsquare s^* by $M^*(s^*)$, and similarly to Lemma 5, we have

Lemma 8. *Let $\Pi' \triangleq \{M^*(s^*) \leq 2nPr\{ORTH(G_i)$ passes $s_j^*\}$ holds for any subsquare$\}$. Then Π' will happen with probability 1 when $n \to \infty$.*

According to Lemma 8, the number of data flows that go through a single subsquare is $O(nPr\{ORTH(G_i)$ passes $s_j^*\})$. Since the buffer occupied by a single data flow in one subsquare is no larger than $2r_mT_s$, and there are at least one node in a subsquare w.h.p., the buffer needed by a node on highways is $O(2r_mT_snPr\{ORTH(G_i)$ passes $s_j^*\})$. Since the number of nodes on highways is $\Theta(n)$, the average buffer occupation in the network is also $O(2r_mT_snPr\{ORTH(G_i)$ passes $s_j^*\})$. Hence we can prove (9) in Theorem 2 with (11).

4.3 Throughput Achieved by Strategy π_2

The transmission route in π_2 actually consists of two parts: the part on highways and the part entering or exiting highways, which we will investigate separately.

Firstly we analyze the throughout achieved on the highways. According to [21], the adjacent subsquares on highways can exchange data at the rate of constant order, which

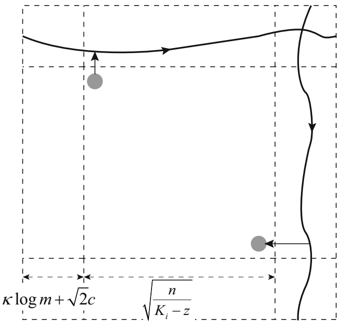


Fig. 3. The method of calculation for $p_i(z)$

is shared by all the data flows going through. Since the number of data flows that go through a single subsquare is $O(nPr\{ORTH(G_i)$ passes $s_j^*\})$, the maximum transmission rate achieved by a data flow on the highways is

$$\Lambda_h(n) \geq \frac{1}{nPr\{ORTH(G_i) \text{ passes } s_j^*\}} = \begin{cases} \Omega(1/\sqrt{n}) & , \quad \alpha \geq 2 \\ \Omega(1/(\sqrt{n} + n^{2-\alpha}\log n)) & , 3/2 < \alpha < 2 \\ \Omega(1/\sqrt{n}\log n) & , \quad \alpha = 3/2 \\ \Omega(1/n^{2-\alpha}\log n) & , 1 < \alpha < 3/2 \\ \Omega(1/n) & , \quad \alpha = 1 \\ \Omega(1/n\log n) & , \quad 0 \leq \alpha < 1 \end{cases} \quad (16)$$

For the throughput for packet entering or exiting highways, we have the following Lemma.

Lemma 9. *The data rate at which packets entering or exiting highways in strategy* π_2 *is* $\Omega(\frac{1}{\sqrt{n}})$.

Proof. In [21], it has been proved that for unicast, since the distance between a node to its corresponding open path is no larger than $\kappa \log m - \varepsilon_m$, according to (1), (2) and (3), the maximum data rate for the one-hop transmission is $\Omega(n^{-c\kappa\gamma/\sqrt{2}}/(\log n)^{2+\alpha})$ $(c\kappa\gamma/\sqrt{2} < 1/2)$. Thus the maximum data rate that can be achieved when packets enter or exit highways is $\Omega(1/\sqrt{n})$. For the case of multicast, because of the fact that a transmission link between a node and its corresponding open path is used by only one data flow, the throughput achieved is same with that in unicast case. Thus we can use the same method in [21] to prove this lemma.

Comparing (16) and Lemma 9, we can see that the bottleneck exists on the highways system. Therefore we have $\Lambda^{(2)}(n) = \Lambda_h(n)$. Thus we have derived (10) and Theorem 2 has been proved.

5 Lower Bound of Buffer Occupation

In this section we investigate the lower bound of buffer occupation in our wireless social network. The result is shown in Theorem 3.

Theorem 3. *The lower bound of buffer occupation is*

$$\begin{cases} \Omega(r_m T_s \sqrt{n}/\sqrt{\log n}) & , \quad \alpha > 3/2 \\ \Omega(r_m T_s \sqrt{n \log n}) & , \quad \alpha = 3/2 \\ \Omega(r_m T_s n^{2-\alpha}/\sqrt{\log n}) & , 1 < \alpha < 3/2 \\ \Omega(r_m T_s n/(\log n)^{3/2}) & , \quad \alpha = 1 \\ \Omega(r_m T_s n/\sqrt{\log n}) & , \quad 0 \le \alpha < 1 \end{cases}$$

To prove Theorem 3, we firstly partition the network into cells of constant side length, and prove the following lemma.

Lemma 10. *The number of cells that are crossed by a multicast group with k destination nodes is $\Omega(\sqrt{kn})$.*

Proof. In [20] it has been proved that

$$Pr\{F \ge \nu_1 \nu_3 n \sqrt{kn}\} \ge 1 - 2exp(-\frac{\nu_3^2}{2\nu_2^2} n),$$

where ν_1, ν_2 and ν_3 are constant, and F is the total number of cells that are passed by multicast groups. Therefore $F \ge \nu_1 \nu_3 n \sqrt{kn}$ with high probability. Since the multicast trees are independent from each other, the expected number of cells crossed by a single multicast tree is $\Omega(\sqrt{kn})$. According to Chernoff bound, it is straight forward that the lower bound for the number of cells crossed by a multicast tree is also $\Omega(\sqrt{kn})$.

Since K_i follows power-law degree distribution, the expected number of hops on a multicast tree is

$$E\{H_i\} \sim \frac{1}{G} \sum_{k=1}^{n-1} \frac{\sqrt{kn}}{k^\alpha} \sim \begin{cases} \sqrt{n} & , \quad \alpha > 3/2 \\ \log n \sqrt{n} & , \quad \alpha = 3/2 \\ n^{2-\alpha} & , 1 < \alpha < 3/2 \\ n/\log n & , \quad \alpha = 1 \\ n & , \quad 0 \le \alpha < 1 \end{cases} . \tag{17}$$

The number of cells that are crossed by different multicast trees is an i.i.d. bernoulli random variable. With Chernoff bound we can easily prove that the number of cells crossed a multicast tree is no less than $\frac{1}{2}E\{H_i\}$ w.h.p..

On the other hand, we repartition the network into squares of size length $\omega = o(\sqrt{n})$. Thus the probability for a source node and one of its destination node to be in a same cell is $\frac{\omega^2}{n} \to 0$, $n \to \infty$. Hence the distance between a source node and its destination node is at least $\Theta(\sqrt{n})$ w.h.p.. For the case of multicast, the number of cells crossed by the route which heads for the farthest destination node is $\Omega(\sqrt{n})$. Based on the fact that the delay of this route is no less than the delay in unicast case, the delay to the farthest destination node is $\Omega(\frac{\sqrt{n}}{\sqrt{\log n}} T_s)$ [24]. Hence the buffer occupied on this route is $r_m T_s \Omega(\frac{\sqrt{n}}{\sqrt{\log n}})$. With strategies similar to TMDA, the delay for a single hop is in constant order. Thus there are only constant difference among the buffer occupation of different nodes. Because the number of cells crossed a multicast tree is no less than $\frac{1}{2}E\{H_i\}$ w.h.p., the buffer occupied by the multicast tree is $\frac{E\{H_i\}}{2\sqrt{n}} \cdot r_m T_s \frac{\sqrt{n}}{\sqrt{\log n}}$. Recall that the number of multicast trees is n, and then we have

$$S_r(n) = \Omega(r_m T_s E\{H_i\}\frac{1}{\sqrt{\log n}}) = \begin{cases} \Omega(r_m T_s\sqrt{n}/\sqrt{\log n}) & , \quad \alpha > 3/2 \\ \Omega(r_m T_s\sqrt{n\log n}) & , \quad \alpha = 3/2 \\ \Omega(r_m T_s n^{2-\alpha}/\sqrt{\log n}) & , 1 < \alpha < 3/2 \\ \Omega(r_m T_s n/(\log n)^{3/2}) & , \quad \alpha = 1 \\ \Omega(r_m T_s n/\sqrt{\log n}) & , \ 0 \le \alpha < 1 \end{cases} . \quad (18)$$

(17) and (18) lead to Theorem 3 directly.

6 Upper Bound of Throughput

We analyze the upper bound of throughput in this section. Firstly we have the following lemma.

Lemma 11. *For a multicast group G_i, if $E\{|EMST(G_i)|\}$ is no less than Q, the throughput of the network satisfies $\Lambda(n) = O(1/Q)$.*

Proof. The number of packets that are transmitted during interval T_c is $r_m T_c n$. Let h_i denote the number of hops during the transmission of packet i, and l_i^h denote the distance at the hth hop. Since $|EMST(G_i)|$ is independent identically distributed and has finite expected value, according to law of large numbers we have

$$\sum_{i=1}^{r_m T_c n}\sum_{h=1}^{h_i} l_i^h \ge r_m T_c \sum_{i=1}^{n} |EMST(G_i)| \ge r_m T_c n Q. \quad (19)$$

With the similar method used in [25], for node n_i, n_j transmitting to n_k, n_l separately, we have $d(n_i,n_j) \ge \Delta(d(n_i,n_k) + d(n_j,n_l))$, where Δ is a constant depending on γ. This result shows that disks of radius Δ times the transmission range centered at the transmitter are disjoint from each other. Since the time taken to transmit a packet is T_s,

$$\sum_{i=1}^{r_m T_c n}\sum_{h=1}^{h_i} \pi(\Delta l_i^h)^2 \le n\frac{T_c}{T_s}. \quad (20)$$

According to Catchy-Schwarz Inequality,

$$[\sum_{i=1}^{r_m T_c n}\sum_{h=1}^{h_i} l_i^h]^2 \le [\sum_{i=1}^{r_m T_c n}\sum_{h=1}^{h_i} (l_i^h)^2][\sum_{i=1}^{r_m T_c n}\sum_{h=1}^{h_i} 1]. \quad (21)$$

Since there are at most n nodes transmitting data in the network,

$$\sum_{i=1}^{r_m T_c n}\sum_{h=1}^{h_i} 1 \le n\frac{T_C}{T_s}. \quad (22)$$

Combining (19), (20), (21) and (22), we have

$$\frac{nT_c}{\pi\Delta^2 T_s} \ge \sum_{i=1}^{r_m T_c n}\sum_{h=1}^{h_i} (l_i^h)^2 \ge [\sum_{i=1}^{r_m T_c n}\sum_{h=1}^{h_i} l_i^h]^2 \frac{T_s}{nT_c} \ge r_m^2 T_c T_c n Q^2.$$

Thus we can get $r_m \le 1/\sqrt{\pi}\Delta T_s Q$, which leads to the result $\Lambda(n) = O(1/Q)$.

Recall that (6) has shown us the order of $E\{|EMST(G_i)|\}$, with Lemma 11 we have

$$\Lambda(n) = \begin{cases} O(1/\sqrt{n}) & , \quad \alpha > 3/2 \\ O(1/\log n\sqrt{n}) & , \quad \alpha = 3/2 \\ O(n^{\alpha-2}) & , 1 < \alpha < 3/2 \\ O(\log n/n) & , \quad \alpha = 1 \\ O(1/n) & , \ 0 \le \alpha < 1 \end{cases} .$$

7 Discussion

7.1 Buffer Occupation Considering Throughput

In the previous sections we have investigated the buffer occupation without regard to r_m. Considering the result in Section 6, we can find that both the lower bounds of throughput achieved by π_1 and π_2 have almost reached their upper bound. Thus to analyze the effect of limited throughput on buffer occupation, we loosely assume that r_m is equal to the lower bound of throughput, which leads to the following results:

$$S^{(1)}(n) = O(1/\log n) \ , \ S^{(2)}(n) = O(1).$$

The results show that the buffer occupied at each node does not increase when the number of nodes become larger, which is opposite to our intuitive understanding. The reason is that the rate of descent for throughput is no less than the climbing speed of buffer occupation. Note that this result is based on the assumption that r_m does not exceed the lower bound of throughput achieved by the transmission strategies.

7.2 Trade-Off between Buffer Occupation and Throughput

Applying strategy π_1, the buffer occupation reaches its lower bound, that is

$$S_r^{(1)}(n) = \begin{cases} \Theta(r_m T_s \sqrt{n}/\sqrt{\log n}) & , \quad \alpha > 3/2 \\ \Theta(r_m T_s \sqrt{n \log n}) & , \quad \alpha = 3/2 \\ \Theta(r_m T_s n^{2-\alpha}/\sqrt{\log n}) & , 1 < \alpha < 3/2 \\ \Theta(r_m T_s n/(\log n)^{3/2}) & , \quad \alpha = 1 \\ \Theta(r_m T_s n/\sqrt{\log n}) & , \ 0 \le \alpha < 1 \end{cases} .$$

However, there exists a gap of $\Theta(\frac{1}{\sqrt{\log n}})$ between $\Lambda^{(1)}(n)$ and its upper bound. For strategy π_2, the throughput achieved reach its upper bound in two intervals:

$$\Lambda^{(2)}(n) = \begin{cases} \Theta(1/\sqrt{n}) & , \ \alpha \ge 2 \\ \Theta(1/\sqrt{n}\log n) & , \alpha = 3/2 \end{cases} .$$

However the buffer occupation is larger than its lower bound by $\Theta(1/(\log n)^{3/2})$ to $\Theta(1/\sqrt{\log n})$. Hence it can be seen that there exists trade-off between buffer occupation and throughput in social wireless network: increasing throughput leads to larger buffer size, and vice-versa. If we consider the case $\alpha \ge 2$, we have $S_r(n) = \Theta(r_m T_s n \Lambda(n))$ and $S(n) = \Theta(T_s n[\Lambda(n)]^2)$.

8 Conclusion

In this paper we investigate buffer occupation in wireless social network. In our network the number of friend nodes for a source node follows power-law degree distribution, and two transmission strategies are proposed to improve the performance of buffer occupation or transmission throughput. Both the upper bound and lower bound for buffer and throughput have been investigated, and the results show that there exists trade-off between them. We also prove that buffer occupation for a single node in wireless social networks does not increase with the increscent number of users, which is another important contribution we have made.

Acknowledgment. This paper is supported by National Fundamental Research Grant (No. 2011CB302701); NSF China (No. 61271219,61202373); Shanghai Basic Research Key Project (No. 11JC1405100, 12JC1405200); China Ministry of Education Fok Ying Tung Fund (No. 122002); China Ministry of Education New Century Excellent Talent (No. NCET-10-0580); China Postdoctoral Science Foundation Grant 2011M500774 and 2012T50417, and SEU SKL project #2012D13.

References

1. Li, J., Blake, C., De Couto, D.S.J., Lee, H.I., Morris, R.: Capacity of ad hoc wireless networks. In: International Conference on Mobile Computing and Networking, pp. 61–69 (2001)
2. Azimdoost, B., Sadjadpour, H.R., Garcia Luna-Aceves, J.J.: Capacity of composite networks: combining social and wireless ad hoc networks. In: Proc. 2011 IEEE Wireless Communications and Networking Conference, pp. 464–468 (2011)
3. Ning, T., Yang, Z., Wu, H., Han, Z.: Self-Interest-Drive Incentives for Ad Dissemination in Autonomous Mobile Social Networks. In: IEEE International Conference on Computer Communications, INFOCOM, Turin, Italy (April 2013)
4. Niyato, D., Han, Z., Saad, W., Hjorungnes, A.: A Controlled Coalitional Game for Wireless Connection Sharing and Bandwidth Allocation in Mobile Social Networks. In: IEEE Globe Communication Conference, Miami, FL (November-December 2010)
5. Zhang, B., Xing, K., Cheng, X., Huang, L., Bie, R.: Traffic Clustering and Online Traffic Prediction in Vehicle Networks: A Social Influence Perspective. In: IEEE INFOCOM, March 25-30, pp. 495–503 (2012)
6. Quercia, D., Lathia, N., Calabrese, F., Di Lorenzo, G., Crowcroft, J.: Recommending social events from mobile phone location data. In: IEEE 10th Int. Data Mining (ICDM) Conf., 2010, pp. 971–976 (2010)
7. Lampos, V., Cristianini, N.: Tracking the flu pandemic by monitoring the social web. In: Proc. 2nd Int. Cognitive Information Processing (CIP) Workshop, pp. 411–416 (2010)
8. Wen, H., Liu, J., Lin, C., Li, P., Fang, Y., Ren, F.: A Storage-friendly Routing Scheme in Intermittently Connected Mobile Network. IEEE Transactions on Vehicular Technology 60(3), 1138–1149 (2011)
9. Herdtner, J.D., Chong, E.: Throughput-Storage Tradeoff in Ad Hoc Networks. In: IEEE INFOCOM (2005)
10. Bodas, S., Shakkottai, S., Ying, L., Srikant, R.: Scheduling in Multi-Channel Wireless Networks: Rate Function Optimality in the Small-Buffer Regime. In: Proceedings of the ACM SIGMETRICS/Performance Conference (June 2009)

11. Wang, X., Yu, T., Xu, Y.: Lower Bound for Node Buffer Size in Intermittently Connected Wireless Networks. IEEE Transactions on Parallel and Distributed Systems 99(1) (2012)
12. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: ACM IMC, New York, NY, USA, pp. 29–42 (2007)
13. Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., Jeong, H.: Analysis of topological characteristics of huge online social networking services. In: Proc. of ACM WWW, New York, NY, USA, pp. 835–844 (2007)
14. Yan, Y., Huang, J., Wang, J.: Dynamic Bargaining for Relay-Based Cooperative Spectrum Sharing. IEEE Journal on Selected Areas in Communications 31(8), 1–14 (2013)
15. Wu, C., Mohsenian-Rad, A.H., Huang, J.: Vehicle-to-Aggregator Interaction Game. IEEE Transactions on Smart Grid 3(1), 434–442 (2012)
16. Zhu, X., Li, P., Fang, Y., Wang, Y.: Throughput and Delay in Cooperative Wireless Networks With Partial Infrastructure. IEEE Transactions on Vehicular Technology 58(8), 4620–4627 (2009)
17. Zhang, B., Cheng, X., Bie, R., Chen, D.: A Community Based Vaccination Strategy Over Mobile Phone Records. In: ACM MHealthSys. (2012)
18. Xu, Y., Wang, X.: Fundamental Lower Bound for Node Buffer Size in Intermittently Connected Wireless Networks. In: IEEE INFOCOM, Shanghai, China, pp. 972–980 (2011)
19. Stuijk, S., Geilen, M., Basten, T.: Exploring trade-offs in buffer requirements and throughput constraints for synchronous dataflow graphs. In: Proceedings of the 43rd Annual Design Automation Conference, pp. 899–904 (2006)
20. Wang, C., Li, X., Jiang, C., Tang, S., Liu, Y., Zhao, J.: Scaling Laws on Multicast Capacity of Large Scale Wireless Networks. In: IEEE INFOCOM, pp. 1863–1871 (2009)
21. Franceschetti, M., Dousse, O., Tse, D.N.C., Thiran, P.: Closing the Gap in the Capacity of Wireless Networks Via Percolation Theory. IEEE Transactions on Information Theory 53, 1009–1018 (2007)
22. Grimmett, G.R.: Percolaiton. Springer (1999)
23. Penrose, M.: Random Geometric Graphs. Oxford University Press, New York (2003)
24. El Gammal, A., Mammen, J., Prabhakar, B., Shah, D.: Throughput-Delay Trade-off in Wireless Networks. In: Proc. IEEE INFOCOM, Hong Kong, pp. 464–475 (2004)
25. Xue, F., Kumar, P.R.: Scaling laws for ad hoc wireless networks: an information theoretic approach. Found. Trends Netw. 1(2), 145–270 (2006)

Efficient Identity-Based Encryption without Pairings and Key Escrow for Mobile Devices

Yan Zhu[1], Di Ma[2], Shanbiao Wang[3], and Rongquan Feng[3]

[1] School of Computer and Communication Engineering,
University of Science and Technology Beijing, 100083, China
zhuyan@ustb.edu.cn

[2] Department of Computer and Information Science,
University of Michigan-Dearborn, 48128, USA
dmadma@umich.edu

[3] School of Mathematical Sciences,
Peking University, Beijing, 100871, China
{wangsb,fengrq}@pku.edu.cn

Abstract. We propose a new construction of identity-based encryption without key escrow over the tradition cryptosystems. The security of our scheme follows from the decisional Diffie-Hellman assumption and the difficulty of a new problem – modular inversion hidden number problem with error (MIHNPwE). The latter can be seen as a generalization of the modular inversion hidden number problem. We give an analysis on the hardness of MIHNPwE by lattice techniques. In our construction, we generate each user's partial private key in the form of an MIHNPwE instance. The hardness of MIHNPwE provides our scheme with resistance against key-collusion attacks from any number of traitors.

1 Introduction

Smart mobile devices have experimented exponential growth over the last several years, but mobile network security has become a major concern to mobile clients throughout the world. In order to solve this issue, more and more cryptographic techniques have been used to enhance the security of these mobile devices. Identity-based encryption (IBE) [1] is also such a technology that can replace the traditional PKI-based public-key cryptosystem because it does not need access certificate authority (CA). In the paradigm of IBE, a sender, knowing the recipient's identity information such as the email address or phone number, can send encrypted messages to the recipient directly by using the recipient's identity information as the public key.

In 1984 Shamir introduced the concept of IBE, but until 2001 Boneh and Franklin proposed the first efficient IBE (BF-IBE) scheme [2]. After that, a lot of IBE systems have been constructed under various assumptions relating to groups with bilinear maps (a.k.a. parings) – a mathematical primitive widely used to build various identity-based cryptographic schemes in existing literature – on elliptic curves. The rich structure of bilinear maps enables various extensions such

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 42–53, 2013.

as Hierarchical IBE, anonymous IBE, and many others. Although pairing-based IBE systems are prevalent, they are not engineering-friendly as implementation of them is much more complicated and often requires the developers have a deeper understanding of the underlying mathematics than implementation of traditional cryptosystems, such as RSA or ElGamal. This usually requires only professionals with appropriate mathematical background can implement them and leads to complex coding which is not easy to understand. Also, their efficiency is only in the asymptotic sense since pairing is usually much slower than modular exponentiation – the major operation – in RSA and ElGamal systems.

To solve this problem, several IBE systems under standard assumptions have been proposed, including the Cocks IBE (Cocks-IBE) scheme [3] and its variants [4,5] under the quadratic residuosity (QR) assumption. Unfortunately, these schemes are considered inefficient as encryption is done on the bit level. This motivates the construction of the BGH-IBE scheme [6], also under the QR assumption, with improved space efficiency. But, its space efficiency is achieved by sacrificing computation efficiency with a complexity of being quartic in the security parameter. Actually, it is more computationally expensive than all "standard" IBE (based on elliptic-curve pairings) schemes and public-key encryption schemes. In addition, lattice-based IBE schemes based on Learning with Error (LWE) problem have been proposed [7]. However, *IBE construction based on traditional cryptosystems – RSA and ElGamal – remains an open question.* This question is also particularly important for mobile devices with limited energy.

More importantly, no matter how an IBE is constructed, in the original definition of IBE, the public key generator (PKG) can generate all the private keys of identities in the system. Thus, it can decrypt any ciphertext in the system. This property is called *key escrow* [8,3]. However, a cryptosystem with key escrow property has some serious disadvantages. First, the PKG has to be fully trusted by all entities. Such a trusted party may not exist. Second, the PKG represents a single point of failure for the IBE system. The compromise of the PKG's master key could be disastrous since the exposure of the master secret undermines all the system secrets (e.g., private keys of all users). One approach to handle the escrow problem is to increase the security of the master secret key by exploiting threshold techniques [9,10]. Gentry and Silverberg presented a method in a hierarchical identity-based encryption scheme [11] to restrict the key escrow function in small areas, but the existence of a master secret key is still a serious threat to the system security.

Contributions. To answer the above questions, we present in this paper a new IBE scheme without key escrow over the integers whose security follows from the DDH assumption and the difficulty of a new hidden number problem. Our IBE scheme performs encryption (and so decryption) on the message level the same as the BF-IBE scheme [8]. The encryption and decryption algorithms are just as in the ElGamal encryption scheme, thus operations in our scheme only involve modular exponentiation and multiplication. With proper parameter choice, our implementation shows it can be even more computation efficient than the pairing-friendly elliptic-curve based BF-IBE scheme proposed [8].

In summary, we make the following contributions: 1) **Definition and hardness analysis of MIHNPwE**; and 2) **Construction of IBE without key escrow over the integers**. We construct an IBE scheme without key escrow under the decisional Diffie-Hellman (DDH) assumption with (near) constant cipher size and computation cost. We prove that the proposed IBE scheme is secure against the selective identity attack in the random oracle model.

Organization. The rest of the paper is organized as follows. The preliminaries and the definitions of our IBE scheme are in Section 2 and 3. We introduce the MHINPwE problem and analyze its hardness in Section 4. We present the construction of our IBE scheme in Section 5 and analyze its security in Section 6. The paper concludes in Section 7.

2 Preliminaries

2.1 Decisional Diffie-Hellman Assumption

Definition 1 (DDH Assumption). *Let $\mathbb{G}$ be a (multiplicative) cyclic group and g be a generator of $\mathbb{G}$. We say the Decisional Diffie-Hellman Assumption (DDH) assumption holds in $\mathbb{G}$, if for any probability polynomial time (PPT) algorithm $\mathcal{A}$, we have $|Pr[\mathcal{A}(g, g^x, g^y, g^{xy}) = 1] - Pr[\mathcal{A}(g, g^x, g^y, g^z) = 1]| < \epsilon$, where ϵ is negligible and the probability is taken over the random choice of x, y, z and the random bits used by $\mathcal{A}$.*

There are several kinds of groups in which DDH is believed to be intractable. For instance, DDH is believed to be hold in multiplicative groups of large prime order, or in groups whose order does not have any small prime divisors. A quintessential example is the subgroup of $\mathbb{Z}_p^*$ with order q where $p = aq+1$, and p, q are primes [12].

2.2 Lattice and Minkowski Theorem

A lattice in the d-dimensional Euclidean space $\mathbb{R}^d$ is the set $\mathcal{L}(\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n) = \{\sum_{i=1}^{n} x_i \boldsymbol{b}_i : x_i \in \mathbb{Z}\}$ of all integral combinations of n linearly independent (column) vector $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n \in \mathbb{R}^d$. The integer n and d are called the rank and dimension of the lattice. A lattice can be conveniently represented by a matrix $B = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n]$. The determinant of the lattice $\mathcal{L}$ is defined as $\det(\mathcal{L}(B)) = \sqrt{\det(BB^T)}$. The most famous problem on lattices is the shortest vector problem (SVP): given a basis of a lattice $\mathcal{L}$, finding a vector $\boldsymbol{u} \in \mathcal{L}$, such that $\|\boldsymbol{v}\| \geq \|\boldsymbol{u}\|$ for any vector $\boldsymbol{v} \in \mathcal{L} \setminus \{\boldsymbol{0}\}$, where the notion $\|\cdot\|$ represents the l_2 norm. The following theorem gives an upper bound on the length of the shortest vector in lattice $\mathcal{L}$ [13].

Theorem 1 (Minkowski). *Any lattice $\mathcal{L}$ of rank n has a non-zero vector $\boldsymbol{v}$ with $\|\boldsymbol{v}\| \leq \sqrt{2n/e\pi} \det(\mathcal{L})^{1/n}$.*

In this paper we use a weaker bound $\sqrt{n} \det(\mathcal{L})^{1/n}$ for simplicity.

3 Definition of IBE without Key Escrow

In the standard definition of IBE, an IBE scheme involves four algorithms: **Setup, Extract, Encrypt** and **Decrypt**. In order to define our IBE without key escrow, we follow the approach from [14] and introduce a new algorithm, denoted by **Publish**, into the standard definition of IBE. This algorithm makes use of the result of **Extract** to generate a real user's public and private key pair (pk_{id}, sk_{id}) into the IBE scheme in order to remove the key escrow property. Based on this approach, we provide our formal definition of IBE scheme without key escrow, which consists of five randomized algorithms, as follows:

- **Setup**(1^{κ}): takes a security parameter κ as input, and generates global public parameters $params$ and a master secret key msk.
- **Extract**$(params, msk, id)$: takes as inputs $params$, msk and an arbitrary public string identity id, and generates a private key d_{id} for id.
- **Publish**$(params, id)$: takes as inputs the public parameters $params$ and an arbitrary id, it outputs a key pair (pk_{id}, sk_{id}) corresponding to the public string id.
- **Encrypt**$(params, id, pk_{id}, M)$: takes as input $params$, a given public key pk_{id} corresponding to the public string id, and generates a ciphertext c_{id} for a message M.
- **Decrypt**$(params, d_{id}, sk_{id}, c_{id})$: takes as input $params$, a ciphertext c_{id}, and a valid private key sk_{id} corresponding to id, and decrypts the ciphertext to get a plaintext M.

We denote the scheme by $\mathcal{E} = (Setup, Extract, Publish, Encrypt, Decrypt)$. The algorithms must satisfy the standard correctness constraint. That is, if d_{id} is the partial private key generated by **Extract** and (pk_{id}, sk_{id}) is generated by **Publish** for a given id as input, then for any message M in message space, let the corresponding ciphertext be $c_{id} = \textbf{Encrypt}(param, id, pk_{id}, M)$, the equation $\textbf{Decrypt}(params, sk_{id}, c_{id}) = M$ holds with all but negligible probability.

3.1 Security Definition

There are several kinds of security notions for identity based encryption schemes. The first formal security definition of IBE was given by Boneh and Franklin [2]. Canetti et al. [15] define a weaker notion of security called selective identity, chosen plaintext secure IBE (IND-sID-CPA). The situation in an IBE scheme without key escrow is a little more complex compared to the basic IBE scheme. Here we consider two types of adversary, which correspond to an adversary with and without the master secret key respectively.

We say an adversary $\mathcal{A}_{\mathrm{I}}$ is a Type-I adversary, if he does not have access to the master secret key, but he has access to the user's private keys without any constraints. We call the owners of these corrupted keys as the traitors. The adversary's objective is to break the ciphertexts, which are sent to an uncorrupted user, by colluding with the traitors. Hence, the Type-I adversary may be related with the identity property, which lies on the choice of the uncorrupted and corrupted users. So, when this attack is described as a security game, we require that $\mathcal{A}_{\mathrm{I}}$ launches a selective identity attack, in which $\mathcal{A}_{\mathrm{I}}$ should output an identity id^* which it wishes to be challenged at the begin of game, and then $\mathcal{A}_{\mathrm{I}}$ makes the following queries (Type-I queries) interacting with a challenger $\mathcal{C}$:

1. For any public identity $id \neq id^*$, $\mathcal{A}_{\mathrm{I}}$ is allowed to query the corresponding partial private key d_{id}.
2. For any public identity id, $\mathcal{A}_{\mathrm{I}}$ is allowed to query the user's public key pk_{id}.
3. For any public identity id, $\mathcal{A}_{\mathrm{I}}$ is allowed to query the user's private key sk_{id}.

We say an adversary is a Type-II adversary, denoted by $\mathcal{A}_{\mathrm{II}}$, if he does have access to the master secret key *msk* but he does not have the user's private key. We will see that the Type-II adversary is related with the key escrow property. Since $\mathcal{A}_{\mathrm{II}}$ knows the master secret key, he can compute the partial private key d_{id} for any public string id by himself. In this game, $\mathcal{A}_{\mathrm{II}}$ launches a selective identity attack when interacting with a challenger $\mathcal{C}$ by making the following queries (Type-II queries):

1. For any public identity id, $\mathcal{A}_{\mathrm{II}}$ is allowed to query the user's public key pk_{id}.
2. For any public identity $id \neq id^*$, $\mathcal{A}_{\mathrm{II}}$ is allowed to query the private key sk_{id}.

Now we define the security for IBE schemes without key escrow. First of all, in terms of the above-mentioned Type-I and Type-II adversaries, we describe the security model by modifying the IND-sID-CPA model from IBE scheme in [16]. In this model $\mathcal{A}_{\mathrm{I}}$ and $\mathcal{A}_{\mathrm{II}}$ are integrated into one adversary $\mathcal{A}$, as follows:

Initial: The adversary $\mathcal{A}$ chooses an identity id^* which it wishes to be challenged and sends it to the challenger $\mathcal{C}$. At the same time, $\mathcal{A}$ declares an attack type to $\mathcal{C}$.

Setup: The challenger $\mathcal{C}$ runs the **Setup** algorithm of the encryption scheme. It gives $\mathcal{A}$ the resulting system parameters *params*. If $\mathcal{A}$ is of Type-I, then the challenger keeps master secret key *msk* to itself, otherwise, it gives *msk* to $\mathcal{A}$.

Phase 1: $\mathcal{A}$ issues a sequence of queries according to the type of the adversary. If $\mathcal{A}$ is of Type-I, he can only make Type-I queries, otherwise he makes Type-II queries. The challenger responds by running the corresponding algorithms according to each query. $\mathcal{C}$ sends the answers to $\mathcal{A}$. These queries may be asked adaptively.

Challenge: Once $\mathcal{A}$ decides that Phase 1 is over it outputs two equal length plaintexts $M_0, M_1 \in \mathcal{M}$ on which it wishes to be challenged. $\mathcal{C}$ picks a random bit $b \in \{0, 1\}$ and computes $c_{id^*} = \textbf{Encrypt}(params, id^*, pk_{id^*}, M_b)$. It sends c_{id^*} as the challenge ciphertext to $\mathcal{A}$.

Phase 2: $\mathcal{A}$ issues additional queries as **Phase 1**.

Guess: Finally, the adversary outputs a guess $b' \in \{0, 1\}$. The adversary wins if $b = b'$.

We refer to such an adversary $\mathcal{A}$ as an IND-sID-CPA adversary. Next, let $\mathcal{O}$ denotes the oracles for some special queries for the user's public keys, the user's private keys, the partial private keys or the hash values. Similarly, let $LR(\cdot,\cdot,b)$ be a left-or-right random Oracle for a random bit $b \in \{0,1\}$, where the output of this function is a value of the first input or the second input according to b. The advantage of the adversary $\mathcal{A}$ in attacking the scheme $\mathcal{E}$ is defined as

$$Adv_{\mathcal{E},\mathcal{A}}^{\text{IND-sID-CPA}}(1^{\kappa}) = |\Pr[\mathcal{A}^{\mathcal{O},LR(\cdot,\cdot,b)}(c_{id^*}) = b] - \tfrac{1}{2}| = |\Pr[b = b'] - \tfrac{1}{2}|.$$

Definition 2. *We say that an IBE system $\mathcal{E}$ is $(t, q_{\mathcal{O}}, \epsilon)$-selective identity, adaptive chosen plaintext secure if for any t-time IND-sID-CPA adversary $\mathcal{A}$ (including $\mathcal{A}_I$ and $\mathcal{A}_{II}$) that makes at most $q_{\mathcal{O}}$ queries (the number of all oracle queries), we have $Adv_{\mathcal{E},\mathcal{A}}^{IND\text{-}sID\text{-}CPA}(1^{\kappa}) < \epsilon$.*

4 MIHNP with Error

In our IBE construction, we define the user's partial private key in a new form which we call the Modular Inversion Hidden Number Problem with Error (MIHNPwE). MIHNPwE is a general instantiation of the modular inversion hidden number problem (MIHNP, [17]) where we use a fixed length random error to replace the original approximation function. In this section we give the definition and hardness analysis of MIHNPwE with two secrets. This will provide a good understanding of the security of the user's private-keys in our IBE scheme against collusion attacks.

First of all, we present the original definition of MIHNP in [17]. Let $\text{MSB}_k(y)$ denote any approximation to y that matches y on the k most significant bits.

Definition 3 (δ-MIHNP). *Let p be a given fixed m-bit prime and k, t be positive integers. Let α be a random hidden element of $\mathbb{Z}_p$. Given p, k, and $(t+1)$ tuples:*

$$\{\langle x_0, \text{MSB}_k(\tfrac{1}{\alpha+x_0} \mod p)\rangle \cdots \langle x_t, \text{MSB}_k(\tfrac{1}{\alpha+x_t} \mod p)\rangle\}$$

for random values $x_0, \cdots, x_t$. The problem is to find α. The δ-MIHNP assumption states that there is no polynomial time algorithm for the MIHNP problem whenever $k < \delta m$, where $0 < \delta < 1$.

In [17], the authors conjecture that the δ-MIHNP assumption holds for any $\delta < \frac{1}{3}$ by using the lattice reduction techniques. Based on their work, we give a variation of the original MIHNP, called MIHNPwE, where we use fixed-size error as the approximation function. Our MIHNPwE is built on two-secret MIHNP and a (prime or composite) modulus n. We give its definition as follows:

Definition 4 (δ-MIHNPwE). *Let n be a fixed m-bit integer and k, t be positive integers. $\alpha \in \mathbb{Z}_n^*$ and $\beta \in \mathbb{Z}_n^*$ are two random hidden elements. Let $s \in \mathbb{N}$, $s \leq k$, $\omega = 2^s$ and $\omega \not| n$. Given n, k, s, and $(t+1)$ tuples:*

$$\{\langle x_i, (\frac{\beta}{\alpha + x_i} \mod n) + \omega \cdot \varepsilon_i\rangle | i = 0, \cdots, t\}$$

where ε_i is a $(m-k)$-bit random integer for all $i \in [0,t]$ and the inverse $(\alpha + x_i)^{-1} \pmod{n}$ exists for all $x_0, \cdots, x_t \xleftarrow{R} \mathbb{Z}$ ($\alpha + x_i$ and n are coprime). The MIHNPwE problem is to find α and β. The δ-MIHNPwE assumption states that there is no polynomial time algorithm solving the MIHNPwE problem whenever both $k < \delta m$, where $0 < \delta < 1$.

In MIHNPwE, the variant ε_i in the above definition is considered as a random error or noise introduced to $\frac{\beta}{\alpha+x_i} \pmod{n}$. In addition, the purpose of the integer n instead of the prime p is to increase the availability and flexibility of MIHNPwE. The integer $\omega = 2^s$ is called a public "shift variant" which shifts the error into a specific position. Actually, when $s = 0$, the error ε_i is introduced to the $(m-k)$ least significant bits; when $s = k$, the error is introduced to the $(m-k)$ most significant bits; when $0 < s < k$, the error is introduced to the middle $(m-k)$ bits. The case when $s = 0$ is equivalent to the $\mathrm{MSB}_k(\cdot)$ function. Hence, MIHNPwE is considered as a more general instantiation of MIHNP. In particular, when $s = 0$ and $k = 0$, the noise ε_i can cover the whole protected value. We call it the full MIHNPwE, which is the most secure form of situation.

Next, we show how to use lattice techniques to analyze the hardness of this problem just as in [17]:

Theorem 2 (Hardness of MIHNPwE). *There exists an effective algorithm to solve the two-secret δ-MIHNPwE when $k \geq \frac{4}{5}m$.*

Theorem 2 shows that 4/5 may be a possible bound for δ-MIHNPwE under general lattice reduction, which means that more than $m/5$ disturbed bits can guarantee the confidentiality of (α, β). In [17], the authors apply a technique due to Coppersmith [18] to make better use of the relations. Namely, instead of using only these relations, they also make use of the relations that are derived by taking products of them. In this way, from the analysis in [17], the authors gave a conjecture that the δ-MIHNP assumption with r-secret holds whenever $\delta < r/(r+2)$. Hence, in the case of two-secret or $r = 2$, this means that the bound of $4m/5$ can be improved to $m/2$ for two-secret MIHNP. Due to the relation that between MIHNPwE is a generalization of MIHNP, we also accept this conjecture.

5 The Proposed Scheme

In this section, we provide a new construction of IBE system without key escrow over the integers. It is designed in multiplicative group of integers modulo p, therefore it is easy to understand and implement. In this scheme, the cryptographic parameters are generated as follows:

Parameter Generator. We say that a randomized algorithm $\mathcal{G}$ is a parameter generator if (1) $\mathcal{G}$ takes a security parameter $k \in \mathbb{Z}^+$, (2) $\mathcal{G}$ runs in polynomial time in k, (3) $\mathcal{G}$ outputs a sufficient large prime number p such that $m > 2k$, where m is the bit length of p, (4) chooses a (multiplicative) cyclic group $\mathbb{G}$ of order p such that the DDH assumption holds, as well as a group generator $g \in \mathbb{G}$,

(5) assigns a random function $H_1 : \{0,1\}^* \rightarrow \mathbb{Z}_p$ and a random function $H_2 : \mathbb{G} \rightarrow \{0,1\}^l$ for some integer l, where $l \in \mathbb{Z}^+$ denotes the length of messages. So, we denote the output of $\mathcal{G}$ by $\mathcal{G}(1^k) = \langle H_1, H_2, p, g, \mathbb{G} \rangle$. The security parameter k is used to determine the size of p. Let the message space be $\mathcal{M} = \{0,1\}^l$ and the ciphertext space be $\mathcal{C} \subseteq \mathbb{G} \times \mathbb{G} \times \{0,1\}^l$.

Setup(1^k)**:** Given a security parameter $k \in \mathbb{Z}^+$, the algorithm works as follows:

Step 1: Run $\mathcal{G}$ on input k to get $\langle H_1, H_2, p, g, \mathbb{G} \rangle$.

Step 2: Choose two integers $\alpha, \beta \xleftarrow{R} \mathbb{Z}_p$ and $\omega = 2^s$ for $s \leq k$.

Step 3: Compute $A = g^\alpha \in \mathbb{G}$ and $B = g^\beta \in \mathbb{G}$.

Finally, this algorithm outputs $params = (H_1, H_2, A, B, g, \omega, \mathbb{G}, p)$ as the system public parameters and the master secret key $msk = (\alpha, \beta)$.

Extract $(params, msk, id)$**:** On input system public parameters $params$, the master secret key msk and a user's identity $id \in \{0,1\}^*$, the algorithm works as follows:

Step 1: Compute $h_{id} = H_1(id) \in \mathbb{Z}_p$, choose a uniformly random ε_i (such that the length of ε_i is at least equals to $m - k$).

Step 2: Compute $d_{id} = (\frac{\beta}{\alpha + h_{id}} \mod p) + \omega\varepsilon_i \in \mathbb{Z}$.

Step 3: Compute $E_{id} = g^{-\omega\varepsilon_i(\alpha + h_{id})} \in \mathbb{G}$.

This algorithm outputs the partial private key d_{id} and publishes E_{id}, then E_{id} is added into the public parameters $params$, or the entity can ask for the value of E_{id} when he wants to know.

Publish$(params, id)$**:** On input public parameters $params$ and a user's id, the algorithm works as follows:

Step 1: Choose a uniformly random $u \xleftarrow{R} \mathbb{Z}_p$.

Step 2: Compute $U_{id} = B^u = g^{\beta u} \in \mathbb{G}$, $A_{id} = A^u$ and $g_{id} = g^u \in \mathbb{G}$.

Finally, this algorithm outputs U_{id}, A_{id}, g_{id} and u. Note that this algorithm is designed for users. Let $(pk_{id}, sk_{id}) = (\{U_{id}, A_{id}, g_{id}\}, u)$. After a user has run this algorithm, $pk_{id} = \{U_{id}, A_{id}, g_{id}\}$ should be published after PKG checks $(A_{id})^{\frac{1}{\alpha}} = g_{id} = (U_{id})^{\frac{1}{\beta}}$ while $sk_{id} = u$ is a private key kept by the user itself.

Encrypt$(params, id, pk_{id}, M)$**:** On input public parameters $params$, a message $M \in \{0,1\}^l$, an identity $id \in \{0,1\}^*$ and the corresponding user public key $pk_{id} = \{U_{id}, A_{id}, g_{id}\}$, the algorithm works as follows:

Step 1: Compute $h_{id} = H_1(id)$ and choose a uniformly random $r \xleftarrow{R} \mathbb{Z}_p$.

Step 2: Compute $c_1 = (A_{id} \cdot g_{id}^{h_{id}})^r \in \mathbb{G}$.

Step 3: Compute $c_2 = (E_{id})^r \in \mathbb{G}$.

Step 4: Compute $c_3 = M \oplus H_2((U_{id})^r) \in \{0,1\}^l$.

Finally, it outputs the ciphertext $c_{id} := (c_1, c_2, c_3)$.

Decrypt$(params, d_{id}, sk_{id}, c_{id})$**:** On input public parameters $params$, partial private key d_{id} and user's private sk_{id} for identity id and a ciphertext $c_{id} = (c_1, c_2, c_3)$, the algorithm works as follows:

Step 1: Compute $T = (c_2)^{sk_{id}} \cdot (c_1)^{d_{id}} \in \mathbb{G}$.

Step 2: Compute $M' = c_3 \oplus H_2(T) \in \{0,1\}^l$.

Finally, it outputs M'.

Correctness. During encryption M is bitwise exclusive-ored with the hash of $(U_{id})^r$, while during decryption c_3 is bitwise exclusive-ored with the hash of $(c_2)^{sk_{id}} \cdot (c_1)^{d_{id}}$. These values used during encryption and decryption are the same while $sk_{id} = u$, because

$$\begin{aligned} T = c_2^{sk_{id}} \cdot (c_1)^{d_{id}} &= (E_{id})^{ru} \cdot ((A \cdot g^{h_{id}})^r)^{u(\frac{\beta}{\alpha+h_{id}}+\omega\varepsilon_i)} \\ &= g^{-ru\omega\varepsilon_i(\alpha+h_{id})} \cdot (g^{r(\alpha+h_{id})})^{u(\frac{\beta}{\alpha+h_{id}}+\omega\varepsilon_i)} \\ &= g^{-ru\omega\varepsilon_i(\alpha+h_{id})} \cdot g^{ru\beta+ru\omega\varepsilon_i(\alpha+h_{id})} = g^{\beta ur} = (U_{id})^r. \end{aligned}$$

So that, $M' = c_3 \oplus H_2(T) = M \oplus H_2((U_{id})^r) \oplus H_2((U_{id})^r) = M$.

We stress that the group $\mathbb{G}$ can be chosen in various forms. Here, we give two concrete instances as follows:

1. We can set the group $\mathbb{G}$ to be the subgroup of size p of $\mathbb{Z}_q^*$ where $q = ap + 1$, $p > q^{\frac{1}{10}}$ and p, q are primes. The DDH assumption [12] is believed to be hold in $\mathbb{G}$.
2. Let $N = PQ$, where $P = 2p + 1, Q = 2q + 1$ are two secure prime number, and p, q are two larger prime integers. The order of the multiplicative group $\mathbb{Z}_N^*$ is $\psi(N) = (P-1)(Q-1) = 4pq$, from which we deduce that there exists a subgroup $\mathbb{G}$ with order p contained in $\mathbb{Z}_N^*$. When p is sufficient large, we can presume that DDH assumption holds in $\mathbb{G}$. Note that, although the above argument can be run on the RSA cryptosystem, the security of scheme is not based on the RSA assumption.

6 Security Analysis

The identity-based encryption is a group-oriented cryptosystem with public parameter, in which all user has their own unique public/private key pairs but the entire system only has a unique master secret key. The size of the group may be very large, so that we must ensure that the security of the user's partial private key and master secret key even for the group size may become infinite in theory. In our scheme, each user's partial private key is constructed in the form of an instance in MIHNPwE. Hence, the security of the master secret key and the partial private keys is guaranteed by the hardness of MIHNPwE. Next, we analyze the security of our scheme from two following aspects:

6.1 Semantic Security for Type-I and Type-II Attacks

Our proposed scheme is IND-sID-CPA secure as described in Definition 2. This means that the ciphertext in our construction can provide the plaintext privacy under a selective identity attack (Type-I attack), as well as the plaintext privacy for system manager (Type-II attack). This kind of privacy is based on semantic security, which means that the challenge ciphertext is indistinguishable from a random element in the ciphertext space.

For Type-I adversary, semantic security is guaranteed by the hardness assumption of DDH problem in $\mathbb{Z}_n$, where DDH is believed to hold if the subgroup of k-th residues modulo a prime n, where $(n-1)/k$ is also a large prime (also called a Schnorr group), or the cyclic group of order $(p-1)(q-1)$, where p and q are safe primes and $n = PQ$ is a RSA-type integer. [1] We state the security of our scheme in the following theorem.

Theorem 3. *If there exists a Type-I IND-sID-CPA adversary $\mathcal{A}_{\mathrm{I}}$ with non-negligible advantage against the above IBE scheme, then there exists an adversary $\mathcal{B}$ which can solve the DDH problem with non-negligible advantage in the random oracle model.*

From this theorem we know that our scheme is selective-identity secure. Namely, without knowing the partial private key d_{id}, an adversary cannot decrypt a message encrypted for the user with identity *id* even if the adversary knows the user's private key sk_{id}.

For Type-II adversary, no key escrow is also guaranteed by the hardness assumption of DDH problem in $\mathbb{Z}_n$. In this attack, the adversary can make complicity with the system manager to know the master secret key. However, in contrast with Type-I adversary, this adversary's objective is to decrypt the user's encryption, or to obtain the secret in the user's private key. We state the security of our scheme in the following theorem.

Theorem 4. *If there exists a Type-II IND-sID-CPA adversary $\mathcal{A}_{\mathrm{II}}$ with non-negligible advantage against the above IBE scheme, then there exists an adversary $\mathcal{B}$ which can solve the DDH problem with non-negligible advantage in the random oracle model.*

From this theorem we known that the user's private key sk_{id} is necessary for decrypting the ciphertext. An adversary cannot decrypt a message encrypted for the user with identity *id*, even with the knowledge of the master key *msk*. Therefore, our scheme is an IBE scheme without key escrow.

Furthermore, our scheme also has the following properties.

Remark 1 (Forward secure). The above scheme is forward secure. Our scheme introduces an extra public and private key pair that only the entity ID knows the private key. Hence even if the master secret key of system is leaked, the prior communications sent to entity ID would not be exposed.

Remark 2. It is worth mentioning that the above scheme implies an ordinary IBE scheme (with key escrow). One can remove the algorithm *Publish* and make some changes to the encryption and decryption algorithm to get an IBE scheme over the integers, but with the key escrow property. The security of the resulting scheme is based on the same assumption. We just omit the details here.

We point out, it is possible to convert any CPA-secure scheme into a CCA-secure one in random oracle model by using techniques such as the Fujisaki-Okamoto method [19]. Note that, due to space limitation, the proofs of the above-mentioned theorems will be provided in Cryptology ePrint Archive.

[1] The DDH assumption does not hold in the multiplicative group $\mathbb{Z}_p^*$, where p is prime. This is because given g^a and g^b, one can efficiently compute the Legendre symbol of g^{ab}, giving a successful method to distinguish g^{ab} from a random group element.

6.2 Security of Private-Keys

In IND-sID-CPA security game, the Type-I adversary is allowed to make $q_{d_{id}}$ partial private key queries. In theory, we do not have a bound for the maximum value of $q_{d_{id}}$. Theorem 3 shows that our IBE scheme is a selective-identity semantically secure scheme under the DDH assumption against a Type-I advesary. However, it does not explicitly explain how we achieve the security of the partial private keys when multiple users have many partial private keys. Bellow, we answer this question based on the result of the MIHNPwE assumption in Section 4, which means that our IBE scheme can resist key-collusion attacks. That is, an adversary, knowing one or multiple key pairs from its colluders, is not able to compute the partial private key of an innocent user.

In our IBE construction, we set the partial private key in the form of $d_{id} = (\frac{\beta}{\alpha+h_{id}} \mod p) + \omega\varepsilon_i$, where h_{id} is the hash value of user's public key (identity), and α, β are the system secrets. According to the definition of two-secret δ-MIHNPwE, we have the following theorem:

Theorem 5. *The IBE scheme constructed in Section 5 is secure against the key collusion attack (i.e., avoiding the leakage of the hidden secrets in the partial private keys) if more than $\delta \cdot \lceil \log_2 p \rceil$ bits of errors are introduced into the partial private keys under the conjecture of the bound of the two-secret δ-MIHNPwE.*

Proof. The proof is obvious. Assume that there exits an effective algorithm $\mathcal{A}$ which can extract the secret information in the partial private keys. Given a δ-MIHNPwE problem with more than $\delta \cdot \lceil \log_2 p \rceil$ noise bits, we can use the algorithm $\mathcal{A}$ to get the hidden secrets because the same format exists in both the MIHNPwE instance and the partial private keys. This contradicts the δ-MIHNPwE assumption. Hence, the theorem holds.

In our construction, we set $m = \lceil \log_2 p \rceil > 2k$, hence the length of noise bits is $(m - k)$ which is greater than $m/2$. According to the conjecture that the δ-MIHNPwE assumption holds when $\delta > 1/2$, we can conclude that the private key are safe in our system.

7 Conclusion

We propose an efficient identity-based encryption scheme without key escrow whose security follows from the DDH assumption and the hardness of MIHNPwE. We give an analysis on the hardness of MIHNPwE by lattice techniques. Our prototype implementation of the proposed scheme shows that it can be more computational efficient than the influential elliptic-curve based IBE scheme.

Acknowledgments. The work of Y. Zhu, R. Feng, and S. Wang was supported by the National Natural Science Foundation of China (Project No. 61170264 and No. 10990011) and the National 973 Program (Project No. 2013CB329606).

References

1. Shamir, A.: Identity-based cryptosystems and signature schemes. In: Blakely, G.R., Chaum, D. (eds.) CRYPTO 1984. LNCS, vol. 196, pp. 47–53. Springer, Heidelberg (1985)
2. Boneh, D., Franklin, M.: Identity-based encryption from the weil pairing. In: Kilian, J. (ed.) CRYPTO 2001. LNCS, vol. 2139, pp. 213–229. Springer, Heidelberg (2001)
3. Cocks, C.: An identity based encryption scheme based on quadratic residues. In: Honary, B. (ed.) Cryptography and Coding 2001. LNCS, vol. 2260, pp. 360–363. Springer, Heidelberg (2001)
4. Di Crescenzo, G., Saraswat, V.: Public key encryption with searchable keywords based on jacobi symbols. In: Srinathan, K., Rangan, C.P., Yung, M. (eds.) INDOCRYPT 2007. LNCS, vol. 4859, pp. 282–296. Springer, Heidelberg (2007)
5. Ateniese, G., Gasti, P.: Universally anonymous ibe based on the quadratic residuosity assumption. In: Fischlin, M. (ed.) CT-RSA 2009. LNCS, vol. 5473, pp. 32–47. Springer, Heidelberg (2009)
6. Boneh, D., Gentry, C., Hamburg, M.: Space-efficient identity based encryption without pairings. In: 48th Annual IEEE Symposium on Foundations of Computer Science, FOCS, pp. 647–657 (2007)
7. Cash, D., Hofheinz, D., Kiltz, E., Peikert, C.: Bonsai trees, or how to delegate a lattice basis. In: Gilbert, H. (ed.) EUROCRYPT 2010. LNCS, vol. 6110, pp. 523–552. Springer, Heidelberg (2010)
8. Boneh, D., Franklin, M.K.: Identity-based encryption from the weil pairing. SIAM J. Comput. 32(3), 586–615 (2003)
9. Gemmel, P.: An introduction to threshold cryptography. In: CryptoBytes, a Technical Newsletter of RSA Laboratories, vol. 2(7) (1997)
10. Chen, L., Harrison, K., Soldera, D., Smart, N.P.: Applications of multiple trust authorities in pairing based cryptosystems. In: Davida, G.I., Frankel, Y., Rees, O. (eds.) InfraSec 2002. LNCS, vol. 2437, pp. 260–275. Springer, Heidelberg (2002)
11. Gentry, C., Silverberg, A.: Hierarchical id-based cryptography. In: Zheng, Y. (ed.) ASIACRYPT 2002. LNCS, vol. 2501, pp. 548–566. Springer, Heidelberg (2002)
12. Boneh, D.: The decision diffie-hellman problem. In: Buhler, J.P. (ed.) ANTS 1998. LNCS, vol. 1423, pp. 48–63. Springer, Heidelberg (1998)
13. Micciancio, D.: The geometry of lattice cryptography. In: Aldini, A., Gorrieri, R. (eds.) FOSAD 2011. LNCS, vol. 6858, pp. 185–210. Springer, Heidelberg (2011)
14. Cheng, Z., Comley, R., Vasiu, L.: Remove key escrow from the identity-based encryption system. In: IFIP International Federation for Information Processing, vol. 155, pp. 37–50 (2004)
15. Canetti, R., Halevi, S., Katz, J.: A forward-secure public-key encryption scheme. In: Biham, E. (ed.) EUROCRYPT 2003. LNCS, vol. 2656, pp. 255–271. Springer, Heidelberg (2003)
16. Boneh, D., Boyen, X.: Efficient selective-id secure identity-based encryption without random oracles. In: Cachin, C., Camenisch, J.L. (eds.) EUROCRYPT 2004. LNCS, vol. 3027, pp. 223–238. Springer, Heidelberg (2004)
17. Boneh, D., Halevi, S., Howgrave-Graham, N.: The modular inversion hidden number problem. In: Boyd, C. (ed.) ASIACRYPT 2001. LNCS, vol. 2248, pp. 36–51. Springer, Heidelberg (2001)
18. Coppersmith, D.: Small solutions to polynomial equations, and low exponent rsa vulnerabilities. J. Cryptology 10(4), 233–260 (1997)
19. Fujisaki, E., Okamoto, T.: Secure integration of asymmetric and symmetric encryption schemes. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, p. 537. Springer, Heidelberg (1999)

A Network Forensics System for Information Leak Events

Tao Zou[1,2,3], Mansoor Alam[1], and Min Song[1]

[1] Department of Electrical Engineering and Computer Science, University of Toledo,
Toledo OH 43606, USA
{tao.zou,mansoor.alam2,min.song}@utoledo.edu
[2] Beijing Institute of System Engineering, Beijing 100101, China
[3] Science and Technology Laboratory of Information System Security,
Beijing 100101, China

Abstract. The events of information leak and illegal content propagation often occur on the network. The existing techniques cannot collect sufficient evidences about users' contents to support forensics for these events. A new approach and a system are proposed which apply Chinese word segment and bloom filter to store the digest of users' contents. With this system, investigators can trace back the events that happened months or even years ago without extra cost of hardware storage.

Keywords: network forensics, users' contents, data processing.

1 Introduction

Nowadays organizations are increasingly attacked by information being revealed to unauthorized parties. Such information leaks can cause harm in a variety of ways. Network forensics is the science that deals with capture, recording, and analysis of network traffic for detecting, and investigating different kinds of security events such as intrusions, misuse, information leaks, and illegal content propagation. There are different systems and techniques people used to deal with security problems on network. Intrusion detection systems (IDSs), which usually focus on network activities, can detect and record the evidences of attack actions by using misuse detection and anomaly detection techniques [1]. Network-based data-leak detection techniques use fuzzy fingerprint to detect accidental data leaks due to human errors or application flaws [2]. However, it is still a big challenge to collect sufficient evidence of information leak and illegal content propagation events very efficiently and accurately due to two unpredictabilities in its application. The first unpredictability is that we are unable to predict what content will be transmitted through the network, nor to define in advance whether it is illegal or allowed to be sent. Thus, the system has to capture all users' contents to collect sufficient potential evidences instead of just logging them according to predefined specific keywords or fingerprints. The second unpredictability is that forensics operation behavior is generally executed after an illegal event is discovered, i.e., someone realized that a classified file was exposed

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 54–63, 2013.

on a public website. However, the response time from the event occurrence to the discovery of the event cannot be predicted, which can range from several days up to months, sometimes even years. Therefore, the system must be capable of storing a large volume of data for a prolonged period of time.

Unfortunately, current technique such as IDS and the existing network forensics systems [3,4,5] cannot resolve these problems very well. Intrusion detection systems log packets only when an alert is triggered. However, almost all information leak events cannot be described by the security policy of IDSs. Existing network forensics systems, for example, the famous network forensics system Infinistream, uses a brute force approach to store all full packets directly, but can only keep a week's worth of data with storage ranging from 1.5Tb - 15Tb [6]. Investigators cannot trace back an event, which happened months ago, due to the lack of source of evidence. Therefore, it becomes a key problem to find an efficient data processing and analysis method that can store more events on the given amount of storage for the application of information leak events forensics. This paper proposes a network forensics system based on users' contents and its data processing, which can effectively support a forensics investigation in a much longer period with the same storage hardware cost.

The rest of the paper is organized as follows. We will first introduce the main idea and the system model in Section 2. In Section 3, the data processing is presented. The experiment results are provided in Section 4. Conclusion and future work are given in Section 5.

2 Main Idea and System Model

2.1 Main Idea

The raw network traffic contains all information such as time, IP address, port number, type of protocol, and users' contents; this information requires a vast amount of storage. However, for application of information leaks and illegal content propagation forensics, not all of this information is necessary. Therefore, it is necessary to figure out how a network forensics system works and what kind of information is needed before we design the system framework and its data processing.

The main purpose of the network forensics system is to collect network evidences, which can help host forensics systems locate the suspicious computers. In general, the investigators have already known that the event had happened before they start the forensics investigation. For example, the investigators may have already detected that a certain classified file had been leaked and now what they need to do is to find out how this event happened and who caused the leak. So the most important investigation results for a network forensics system are to provide the answers of whether the event happened in this network, when it happened, and who is the suspicious person. These triple W (Whether, When, and Who) investigation results are sufficient to help host forensics systems to start a further investigation for more reliable evidences.

What a network forensics system needs to capture and store is the information which can help answer the triple W questions, that is much less than the raw network data. Since privacy issues have always been a major concern in computer forensics and in case of any investigation [7], we employ the bloom filter algorithm, which uses a hashed version of the content to represent the original content, to further reduce the data volume as well as to avoid the problem of invasion of network users' privacy without the loss of triple W information.

2.2 System Model

We design and build a Distributed Network Forensics (DNF) system based on users' contents. The system is composed of one client console, several data centers and data engines. The system is illustrated as Figure 1.

The client console receives the investigators' inquiries and decomposes them to multiple queries to different data centers, and returns the results to investigators through the human-machine interface. Data centers are deployed in different network domains to store potential evidence data and help the client console to finish the users' inquires. A data center is composed of three functional modules. The first one is the communication module, which is responsible for receiving the data from the data engines as well as queries from the client console. The second one is the database, which stores the history data, including communication

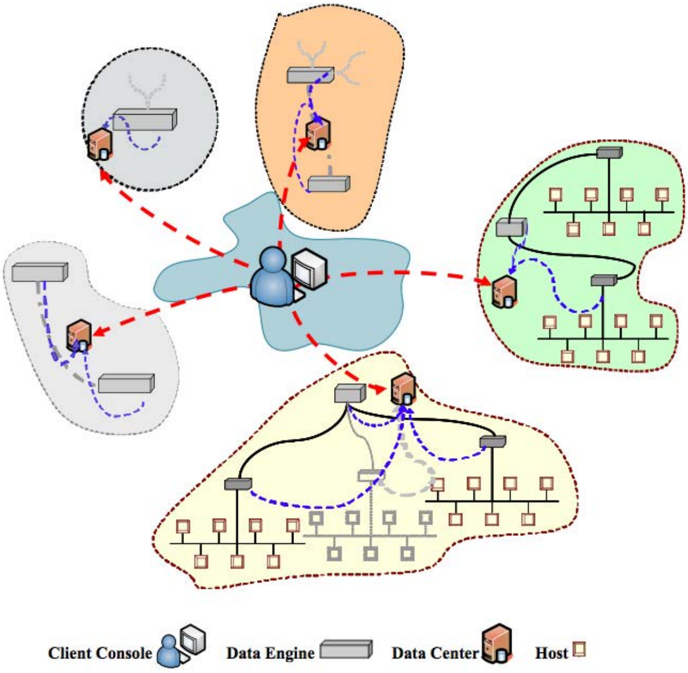


Fig. 1. System model

session records and the digest of the users' contents. The third module is the inquiry and analysis module, where time, IP address records, and content digests are processed to help find the triple W results.

There are one or more data engines in each network domain. Data engine captures the network packets and processes them to the specific format that is suitable for prolonged storage. Each engine is composed of a packet capturing module, a data processing module and a communication module. The packet capturing module captures all the packets, including all users' contents in the network traffic. The data processing module extracts the useful information from the raw network data and records the time and the IP addresses between source and destination hosts. The communication module will upload the data such as session records and the digests of users' contents to data center as potential evidence. Next, we focus on the data processing module.

3 The Data Processing

The data processing is an essential part of the data engines. Figure 2 shows how raw network data are processed at the stage of data capturing and storing. The dashed line denotes the data flow and the solid line denotes the program control flow, and the oval is the data form and the rectangle is the function of the data engine.

The raw network data are the network packets captured by the data engine, which always connected to the switcher's monitoring port. The protocol parser will first analyze them to decide which packets need to be processed further and which should be discarded immediately. For example, if it is an Internet Control Message Protocol (ICMP) protocol packet, that means it only contains network control information, and it is useless for information leaks forensics. So these packets will be discarded. But if it is a File Transfer Protocol (FTP), Hypertext Transfer Protocol (HTTP), or Simple Mail Transfer Protocol (SMTP) protocol packet, it should be saved for further process because this packet may contain users' content.

There are many kinds of content types for different Internet users, e.g. text, image files, audio or video files. Since it is too difficult to parse and understand the content of multimedia files [8], this paper will focus on the document files including TXT/DOC/PPT/PDF. These users' files will be detected and filtered by file parser and filter. The character extractor will then pick out all the text content from different users' files and transform them into pure-text type with normalized character coding. Although these text contents need much less storage space than the raw network packets, it contains almost all evidences needed for information leaks forensics. In order to further reduce the storage space and protect network users' privacy, Chinese word segment and the bloom filter are used to further process those text contents [9].

Other information about communication sessions such as time and IP addresses of hosts are recorded into the session records files at the same time. This information is appended to the target text slices before the bloom filter transforms them into digests. Because time and IP addresses are the key attributions

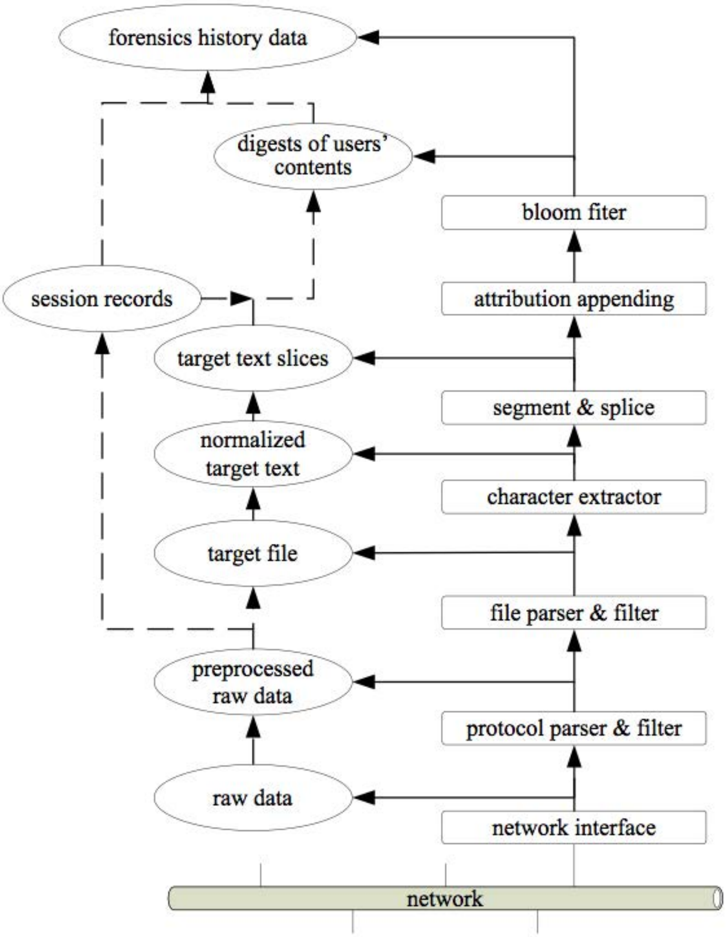
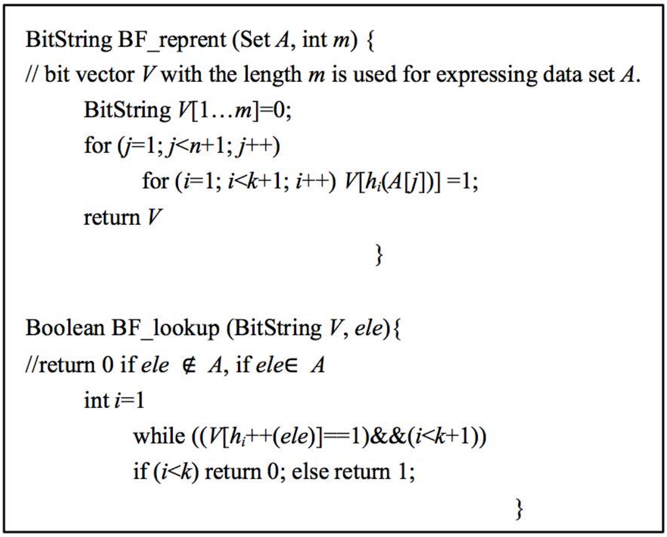


Fig. 2. Data processing

of a session, which contain the information to answer WHEN and WHO, this process is call attribution appending.

A bloom filter is a data structure, which has a strong space advantage over other data structures for representing sets. It can rapidly and memory-efficiently tell whether an element is a member of a set. This data structure can be described as follows: the bit vector V with the length m is used for expressing the element set $A = \{a_1, a_2, \ldots, a_n\}$, h_i are k Hash functions with uniform distribution features, and $\forall x \in A,\ h_i(x) \in \{1, 2, \ldots, m\}$, shown in Figure 3.

Since a bloom filter does not store the original elements, it does not support browsing the original elements either. This feature is very helpful for protecting network users' privacy, because no investigator can pick an element out of a bloom filter to get its content but only can do a membership query to know whether an element is a member of the set or not.



```
BitString BF_reprent (Set $A$, int $m$) {
// bit vector $V$ with the length $m$ is used for expressing data set $A$.
        BitString $V[1\ldots m]=0$;
        for ($j=1; j<n+1; j$++)
                for ($i=1; i<k+1; i$++) $V[h_i(A[j])] =1$;
        return $V$
                                        }

Boolean BF_lookup (BitString $V$, $ele$){
//return 0 if $ele \notin A$, if $ele \in A$
        int $i=1$
              while (($V[h_i$++$(ele)]$==1)&&($i<k+1$))
              if ($i<k$) return 0; else return 1;
                                                    }
```

Fig. 3. Bloom filter algorithm

As for the DNF system, the input set element to a bloom filter is the target text slice appended session attributions. Assuming that the average length of target text slices is g characters (a two-byte Chinese character encoding method is adopted), the storage space of $g \times n \times 16$ bits is needed to directly store a set containing n target text slices. But only the storage space of m bits is needed for a bloom filter. In practice, we have $m \ll g \times n \times 16$.

A bloom filter has a strong space advantage and there is no false negative, but false positive retrieval results exist. The false positive rate can be calculated according to the following formula,

$$FP = (1-(1-\frac{1}{m})^{kn})^k \approx (1-e^{-\frac{kn}{m}})^k \tag{1}$$

It can be derived that by using the bloom filter, the data volume will be reduced by percentage of $\frac{m}{g \times n \times 16}$. When $k = ln2 \times \frac{m}{n}$, FP achieves the minimum value $FP_{min} \approx 0.6185^{\frac{m}{n}}$. The false positive rate, and data reduction rate are concretely analyzed in the following section.

The query granularity of the DNF system depends on the size of the element input to the filter. If we take a file as an element, and the investigators do not catch the whole file (for example only get a paragraph of that file), investigators could not know whether that event happened in the network or not, because the system does not support the membership query of a paragraph. Another important consideration is that in many cases, the file that was discovered by investigators before the investigating is not exactly the same as that transmitted through network before. It may be edited such as entering one more line break or adding some extra words in it. Although these two files are definitely related, the system without a small granularity is unable to tell the relationship between them. A smaller granularity can offer a higher level of flexibility and adaptability. That is why we use Chinese word segment algorithm to cut the whole text file into multiple slices. More information about this issue can be found in [9].

In order to get the evidence of the information leak event, an investigator needs only to do the following three steps.

Step 1: estimating a time range from t_1 to t_2, in which the event might happen, and searching all sessions happened during this period in the session record database. The system will return IP addresses pairs such as $< sourceIP_1, destinationIP_1 >, < sourceIP_2, destinationIP_2 >, \ldots, < sourceIP_m, destinationIP_m >$. These IP addresses pairs can be written as $\{IP_{s_1}, IP_{s_2}, \ldots, IP_{s_m}\}$.

Step 2: inputting the exposed classified content, which the investigator has already known, into the system. The system will process it by segment and splice function module and output target text slices, for example $\{S_1, S_2, S_3, \ldots, S_n\}$. These target text slices will be appended with those IP addresses pairs separately as $\{S_1 + IP_{s_1}, S_2 + IP_{s_1}, S_3 + IP_{s_1}, \ldots, S_n + IP_{s_1}; S_1 + IP_{s_2}, S_2 + IP_{s_2}, S_3 + IP_{s_2}, \ldots, S_n + IP_{s_2}; \ldots\}$. The system will do membership inquires for all $S_i + IP_{s_j}$ (i from 1 to n, j from 1 to m) one by one in the forensics history database.

Step 3: if for a certain j, the membership inquires for all $S_i + IP_{s_j}$ (i from 1 to n) are positive, it means this content was transmitted from *source* IP_j to *destination* IP_j. The time of the session is just the time when the information leak event happened. Otherwise, the investigator needs to go to step 1 to estimate a new time period.

4 Numerical Results

4.1 Performance Analysis

According to Eq. (1), the curve of parameter k and m/n at certain false positive rate FP is shown in Figure 4. X axes is k and Y axes is m/n. When eight hash functions are explored, and the system false positive rate is 0.01, m/n is about 10. When n nonrepeated target text slices are stored and assuming that the average length of the target text slice is 8-characters, the ratio of data volume reduction can be about 1/12.8. Considering that repeated text slices are always contained in actual situation and the useless Chinese auxiliary words such as ‘de’, ‘le’, ‘ma’ will be filter out, the storage space efficiency of the system can be improved to about 15 times in practice.

As mentioned previously in Section 3, there are false positive retrieval results in the bloom filter. Since every user’s content file will be sliced into several target text slices in the proposed DNF system and we need one query to know whether a slice is in the set or not; More than one query will be done to inquire whether a whole file is in the set or not. Most often, at least five queries are needed. The final false positive rate of an investigation for a certain information leak event is only $10^{-2\times5} = 10^{-10}$, which is almost negligible.

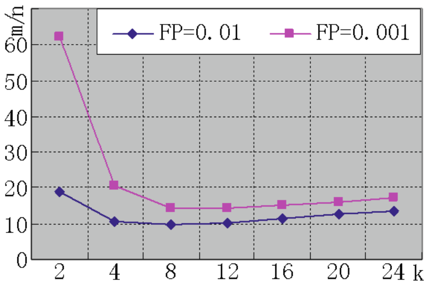
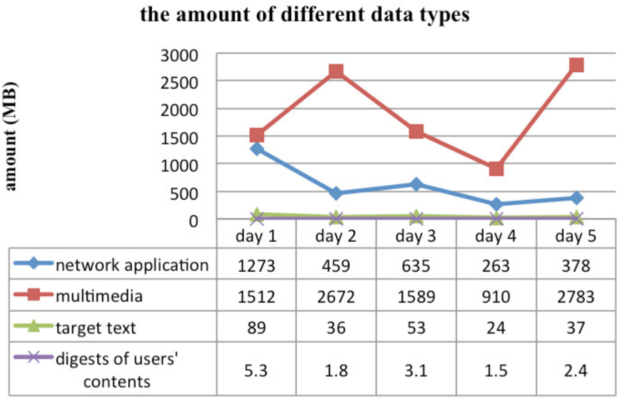


Fig. 4. Curves of m/n versus k

4.2 Experiments Result

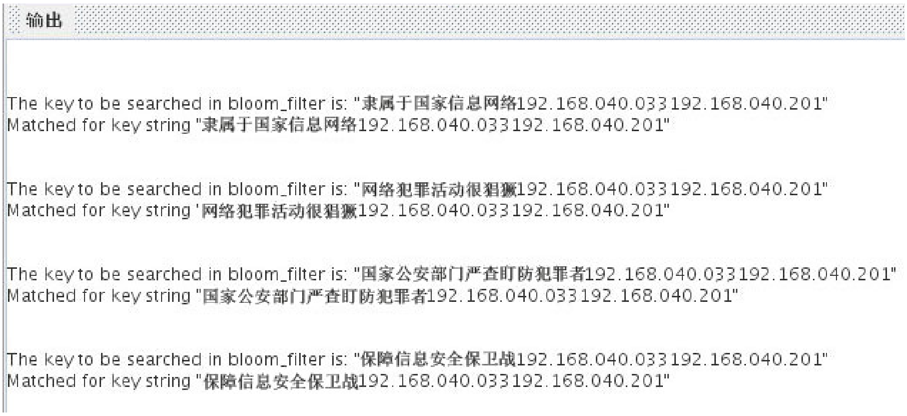
Experiments were done in an office network of a department in BISE (Beijing Institute of System Engineering). There are 30 computers divided into two different domains. The network is used to support office service and thus has no heavy traffic. The system, which includes two data engines and two data centers, was deployed on the network. The data engine was connected to the monitoring port of the switcher to collect all network data. We collected the data in five weekdays. Figure 5 shows the data amount of different data types.



	day 1	day 2	day 3	day 4	day 5
network application	1273	459	635	263	378
multimedia	1512	2672	1589	910	2783
target text	89	36	53	24	37
digests of users' contents	5.3	1.8	3.1	1.5	2.4

Fig. 5. The amount of different data types

The total data including network application, multimedia and target text data for five days is 12713 MB. The volume of the target text data is 239 MB. The amount of the digests of users' contents is 14.1 MB, only about 1/17 of the target text data amount. Thus if a common system, which saves target text directly, can store potential evidence data of events in a month range, our system can



```
The key to be searched in bloom_filter is: "隶属于国家信息网络192.168.040.033192.168.040.201"
Matched for key string "隶属于国家信息网络192.168.040.033192.168.040.201"

The key to be searched in bloom_filter is: "网络犯罪活动很猖獗192.168.040.033192.168.040.201"
Matched for key string '网络犯罪活动很猖獗192.168.040.033192.168.040.201"

The key to be searched in bloom_filter is: "国家公安部门严查盯防犯罪者192.168.040.033192.168.040.201"
Matched for key string "国家公安部门严查盯防犯罪者192.168.040.033192.168.040.201"

The key to be searched in bloom_filter is: "保障信息安全保卫战192.168.040.033192.168.040.201"
Matched for key string "保障信息安全保卫战192.168.040.033192.168.040.201"
```

Fig. 6. The result of forensics

store them in a range of 17 times longer than the common one. That is nearly one and a half years. DNF greatly improved the system's trace back capability.

During the experiment time, a test classified DOC file was transferred between two hosts we chose randomly. The IP addresses are 192.168.40.33 and 192.168.40.201. We began to investigate this information leak event 30 days after it had happened. The administrator searched all sessions happened during these 5 days and input only one paragraph randomly selected from the DOC file into the system. When membership inquires for IP address pair (192.168.40.33 and 192.168.40.201) and the paragraph were processing, the system got the matched results, which are show in Figure 6. The prompt of "Matched for keys string" means the following words marked by the quotation in that line had been transferred between these two hosts in the past. The exact time when that event happened can be searched out through session record database.

5 Conclusion and Future Work

The existing techniques and systems cannot support network forensics for information leak events due to the huge amount of potential evidence data. This paper presents a distributed system and its data processing, which applies Chinese word segment and bloom filter to store the digest of users' contents and can provide WHETHER/WHEN/WHO information for investigators. This new system can not only support to trace back events happened a long time ago, but also help protecting network users' privacy. The experiment results demonstrate the system's storage efficiency and its investigating capability.

Although only text contents are demonstrated and analyzed in this paper, the system can also support other types of users' contents such as images and audios, where each multimedia file needs to be treated as a signal element for the bloom filter. In the future, content parsing and processing for multimedia files will be further investigated to help improve the system's analysis capability for multimedia.

Acknowledgements. The authors would like to thank Xiaohua Xu, Changlong Chen and Zhuang Zhang for their participation in discussions about this research.

References

1. Depren, O., Topallar, M., Anarim, E., Ciliz, M.K.: An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks. Expert Systems with Applications 29, 713–722 (2005)
2. Shu, X., Yao, D.D.: Data leak detection as a service: challenges and solutions. Technical Report TR-12-10, Computer Science, Virginia Tech. (2012)
3. Kaur, J., Singh, G., Singh, M.: Design & Implementation of Linux based Network Forensic System using Honeynet. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 1, 504 (2012)
4. Hunt, R., Zeadally, S.: Network Forensics: An Analysis of Techniques, Tools, and Trends. Journal of Computer 45, 36–43 (2012)
5. Pilli, E.S., Joshi, R.C., Niyogi, R.: Network Forensic Frameworks: Survey and Research Challenges. Digital Investigation 7, 14–27 (2010)
6. Thomas, A.: A Distributed Network Performance and Traffic Analyser. In: Dissertation of Science in Computer Information System. the University of Bath (2009)
7. Aminnezhad, A., Dehghantanha, A., Abdullah, M.T.: A Survey on Privacy Issues in Digital Forensics. International Journal of Cyber-Security and Digital Forensics (IJCSDF) 1, 311–323 (2012)
8. Battiato, S., Emmanuel, S., Ulges, A., Worring, M.: Multimedia in Forensics, Security, and Intelligence. IEEE MultiMedia 19, 17–19 (2012)
9. Zou, T., Xu, B., Huang, M., Liu, L., Zhao, G.: A Method and A Device of Chinese Text Processing. Chinese Patent: CN200910086633.8 (2009)

Maximum Independent Set of Links with a Monotone and Sublinear Power Assignment

Chao Ma, Fahad Al-dhelaan, and Peng-Jun Wan

Illinois Institute of Technology, Chicago, IL 60616

Abstract. This paper studies the problem of selecting a maximum independent set of links with a fixed monotone and sublinear power assignment under the physical interference model. The best-known approximation bound for this problem is a very large constant. In this paper, we present an approximation algorithm for this problem, which not only has a much smaller approximation bound but also produces an independent set of links with a stronger property, i.e., strong independence.

1 Introduction

Consider a multihop wireless network consisting of a set V of nodes with maximum transmission power P. The strength of a transmitted signal attenuates with a path loss factor $\eta r^{-\kappa}$, where r is the distance from the transmitter, κ is *path-loss exponent* (a constant between 2 and 5 depending on the wireless environment), and η is the *reference loss factor*. The signal quality perceived by a receiver is measured by the *signal to interference and noise ratio* (*SINR*), which is the quotient between the power of the wanted signal and the total power of unwanted signals (i.e., interferences) and the ambient noise ξ. In order to correctly interpret the wanted signal, the SINR must exceed certain threshold σ under the physical interference model. Thus, for a communication from a node u to a node v to be possible even without any interference, the transmission power of u should exceed

$$p_0(u,v) := \sigma\xi_0 \|uv\|^{\kappa},$$

where $\|uv\|$ is the Euclidean distance between u and v and $\xi_0 = \frac{\xi}{\eta}$. Thus, the largest possible set of communication links is

$$L = \left\{(u,v) \in V^2 : p_0(u,v) < P, u \neq v\right\}.$$

For each $a = (u,v) \in L$, we use $p_0(a)$ to denote $p_0(u,v)$, and use $\ell(a)$ to denote its length $\|uv\|$. In addition, the distance between the sender of a link $a \in L$ and the receiver of a link $b \in L$ is denoted by $\ell(a,b)$.

Consider a non-empty subset A of L and a power assignment p on A satisfying that for each $a \in A$, $p_0(a) < p(a) \leq P$. A subset I of A is said to be an *independent set* (abbreviated with IS) of A if I is node-disjoint and for each $a \in I$,

$$\frac{p(a)\,\ell(a)^{-\kappa}}{\sum_{b\in I\setminus\{a\}} p(b)\,\ell(b,a)^{-\kappa} + \xi_0} > \sigma.$$

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 64–75, 2013.

The independence can be conveniently characterized in terms of relative interference. For two links a and b in A, the *relative interference* of a toward b, denoted by $RI_p(a,b)$, is defined as follows: If a and b share a common node, then $RI_p(a,b) = \infty$; otherwise,

$$RI_p(a,b) = \sigma \frac{p(a)\,\ell(a,b)^{-\kappa}}{(p(b) - p_0(b))\,\ell(b)^{-\kappa}}.$$

It's easy to verify for any an $I \subseteq A$ is independent if and only if for any $a \in I$,

$$\sum_{b \in I \setminus \{a\}} RI_p(b,a) < 1.$$

The problem of finding a largest IS of B is known as **Maximum Independent Set of Links** (**MISL**).

Constant-approximation algorithms for **MISL** have been developed with various fixed a monotone and sub-linear power assignments [1,2,3,4,5,6,7]. A power assignment p is said to be *monotone* if $p(a)$ is non-decreasing with $\ell(a)$, and to be *sub-linear* if $p(a)\,\ell(a)^{-\kappa}$ is non-increasing with $\ell(a)$. With a general monotone and sub-linear power assignment, the best-known approximation bound for **MISL** is $960 \cdot 3^{\kappa}$ [4]. This paper presents improved approximation algorithms for **MISL** with a fixed monotone and sublinear assignment p. Our algorithms not only output a *strongly* independent set I in the sense that for any $a \in I$,

$$\sum_{b \in I \setminus \{a\}} (RI_p(a,b) + RI_p(b,a)) < 1,$$

but also have a much smaller approximation approximation bound $64\left(2 + (2\sigma)^{-1/\kappa}\right)^{\kappa}$. In addition, we present an extension to the multi-channel multi-radio (MC-MR) wireless networks.

The remaining of this paper proceeds as follows. In Section 2, we introduce a general conflict system and develop a general algorithmic framework. In Section 3, we explore a geometric property of the independence with an arbitrary power assignment. In Section 4, we present an improved approximation algorithm for **MISL** with a fixed monotone and sub-linear power assignment. The following standard terms and notations are adopted throughout this paper. Consider a non-empty set E.

- For any real-valued function f on E and any $S \subseteq E$, $f(S)$ represents $\sum_{e \in S} f(S)$.
- For any real-valued function f on $E \times E$, and any $a \in E$, and any $S \subseteq E$, $f(a,S)$ represents $\sum_{b \in S} f(a,b)$, and $f(S,a)$ represents $\sum_{b \in S} f(b,a)$. In general, for any pair of subsets E_1 and E_2 of E, $f(E_1,E_2)$ represents

$$\textstyle\sum_{e_1 \in E_1} \sum_{e_2 \in E_2} f(e_1,e_2).$$

- Let $\prec$ be an ordering of E. For any pair of links $a, b \in E$, $a \prec b$ means that a appears before b in the ordering $\prec$; $a \preceq b$ represents that either $a \prec b$ or $a = b$. Consider any $a \in E$ and any $B \subseteq A$. $a \prec B$ (respectively, $a \preceq B$) means that $a \prec b$ (respectively, $a \preceq b$) for each $b \in B$. We use $B_{\prec a}$ (respectively, $B_{\preceq a}$) to denote the set of links $b \in B$ satisfying that $b \prec a$ (respectively, $b \preceq a$).

2 Symmetric Conflict System

A symmetric conflict system is specified by a pair (A, ρ) where A is a non-empty finite set and ρ is a symmetric "conflict" function from A^2 to $\mathbb{R}_+ \cup \{\infty\}$. Consider a parameter $\phi > 0$. A subset I of A is said to be a *ϕ-restricted independent set* (IS) of A with respect to ρ if $\rho(I \setminus \{a\}, a) < \phi$ for each $a \in I$. For any ordering $\prec$ of A, a subset J of A is said to be a *ϕ-restricted inductively independent set* (IIS) of A in $\prec$ with respect to ρ if $\rho(J_{\prec a}, a) < \phi$ for each $a \in J$. A ϕ-restricted IIS J of A in $\prec$ can be computed greedily as follows: Initially, J is empty. For each $a \in A$ in the ordering $\prec$, it is added to J if $\rho(J, a) < \phi$. Such J is referred to as the *greedy ϕ-restricted IIS* of A in $\prec$.

2.1 Extraction of an Independent Set

In this subsection, we present a simple algorithm **ExtractIS**(ϕ) which extracts a ϕ-restricted IS I from A for a given parameter $\phi > 0$.

Let ρ_ϕ be the function on A^2 defined by

$$\rho_\phi(a, b) = \min\{\phi, \rho(a, b)\}.$$

for any a and b in A. Then, (A, ρ_ϕ) is also a symmetric conflict system. Clearly, any ϕ-restricted IS of A with respect to ρ is also a ϕ-restricted IS of A with respect to ρ_ϕ, and vice versa. The average conflict of A with respect to ρ_ϕ is defined to be

$$\frac{\sum_{a \in A} \rho_\phi(a, I \setminus \{a\})}{|A|} = \frac{\sum_{a \in A} \sum_{b \in A \setminus \{a\}} \rho_\phi(a, b)}{|A|}.$$

Let φ denote the average conflict of A with respect to ρ_ϕ and let $t = \min\left\{1, \frac{\phi}{\varphi}\right\}$. The algorithm **ExtractIS**(ϕ) is outlined in Table 1. It initializes I to an empty set and another set F to A, and proceeds in two phases:

- **Growing Phase:** While F is non-empty, remove any a from F, and add it to I if
$$\rho(a, I) + t\rho_\phi(a, F) < \phi.$$
- **Pruning Phase:** While I is not a ϕ-restricted IS, remove from I any a satisfying that $\rho(a, I \setminus \{a\}) \geq \phi$.

Table 1. The description of the aslgorithm **ExtractIS**(ϕ)

```
ExtractIS($\phi$):
$F \leftarrow A$, $I \leftarrow \emptyset$, $t \leftarrow \min\left\{1, \frac{\phi}{\varphi}\right\}$; //Initialization
while $F \neq \emptyset$,        //Growing Phase
    remove any $a$ from $F$;
    if $\rho(a, I) + t\rho_\phi(a, F) < \phi$, then $I \leftarrow I \cup \{a\}$;
while $I$ is not a $\phi$-restricted IS w.r.t. $\rho$,        //Pruning Phase
    remove from $I$ any $a$ with $\rho(a, I \setminus \{a\}) \geq \phi$;
return $I$.
```

The theorem below presents lower bounds on $|I|$.

Theorem 1. *If $\phi \geq \varphi$ then $|I| \geq \left(1 - \frac{\varphi}{2\phi}\right)|A|$; otherwise, $|I| > \frac{\phi}{2\varphi}|A|$.*

Proof. We define a quadratic function f from the set of vectors $x \in [0,1]^{|A|}$ indexed by the set of links in A to real numbers by

$$f(x) = \sum_{a \in A} x_a - \frac{1}{2\phi} \sum_{a \in A} \sum_{b \in A \setminus \{a\}} \rho_\phi(a, b)\, x_a x_b.$$

Then for each $a \in A$, f is a linear function of x_a with slope

$$1 - \frac{1}{\phi} \sum_{b \in A \setminus \{a\}} \rho_\phi(a, b)\, x_b,$$

which is referred to as the *a-slope of f at x*. Define a vector $y \in [0,1]^{|A|}$ by $y_a = t$ for each $a \in A$. Then,

$$\begin{aligned} f(y) &= |A|\, t - \left(\frac{1}{2\phi} \sum_{a \in A} \sum_{b \in A \setminus \{a\}} \rho_\phi(a, b)\right) t^2 \\ &= |A|\, t - |A| \frac{\varphi}{2\phi} t^2 \\ &= |A|\, t \left(1 - \frac{\varphi}{2\phi} t\right). \end{aligned}$$

If $\phi \geq \varphi$ then $t = 1$ and $f(y) = \left(1 - \frac{\varphi}{2\phi}\right)|A|$; otherwise, $t = \frac{\phi}{\varphi}$ and $f(y) = \frac{\phi}{2\varphi}|A|$.

We initialize a vector $x \in [0,1]^{|A|}$ to y, and modify it through the executions of the algorithm as follows:

- For each iteration in **Growing Phase**, let a be the link removed from F, and reset x_a to 1 if a is added to I, and to 0 otherwise.
- For each iteration in **Pruning Phase**, let a be the link removed from I, and reset x_a to 0.

We remark that at the end of the **Growing Phase**, $x \in \{0,1\}^{|A|}$. We shall show that $f(x)$ is non-decreasing after each update. As the result, the final output I satisfies that

$$|I| = \sum_{a \in A} x_a \geq f(x) \geq f(y),$$

and hence the theorem holds.

We first show that $f(x)$ is non-decreasing after each iteration in **Growing Phase**. Consider a particular iteration of **Growing Phase** and let a be the link removed from F. At the beginning of this iteration, the a-slope of f at x is

$$\begin{aligned} & 1 - \frac{1}{\phi} \sum_{b \in A \setminus \{a\}} \rho_\phi(a,b)\, x_b \\ & = 1 - \frac{1}{\phi} \left(\sum_{b \in I} \rho_\phi(a,b) + t \sum_{b \in F} \rho_\phi(a,b) \right) \\ & = 1 - \frac{1}{\phi} \left(\rho_\phi(a,I) + t\rho_\phi(a,F) \right). \end{aligned}$$

If

$$\rho(a,I) + t\rho_\phi(a,F) < \phi,$$

then

$$\rho_\phi(a,I) + t\rho_\phi(a,F) < \phi,$$

the a-slope of f at x is positive; so by increasing x_a to 1, $f(x)$ increases. Otherwise,

$$\rho_\phi(a,I) + t\rho_\phi(a,F) \geq \phi,$$

and hence the a-slope of f at x is non-positive; so by decreasing x_a to 0, $f(x)$ either increases or remains unchanged. So, in either case, $f(x)$ is non-decreasing after each iteration in **Growing Phase**.

Next, we show that $f(x)$ is non-decreasing after each update in **Pruning Phase**. Consider a particular iteration of **Pruning Phase** and let a be the link removed from I. At the beginning of this iteration, the a-slope of f at x is

$$\begin{aligned} & 1 - \frac{1}{\phi} \sum_{b \in A \setminus \{a\}} \rho_\phi(a,b)\, x_b \\ & = 1 - \frac{1}{\phi} \sum_{b \in I \setminus \{a\}} \rho_\phi(a,b) \\ & = 1 - \frac{1}{\phi} \rho_\phi(a, I \setminus \{a\}). \end{aligned}$$

Since

$$\rho(a, I \setminus \{a\}) \geq \phi,$$

we have

$$\rho_\phi(a, I \setminus \{a\}) \geq \phi,$$

and hence the a-slope of f at x is non-positive; so by decreasing x_a to 0, $f(x)$ either increases or remains unchanged.

We remark that if $\phi \geq \varphi$, then $t = 1$. In this case, we can simply initialize I to A, skip the **Growing Phase**, and jump to the **Pruning Phase** immediately**:** While I is not a ϕ-restricted IS, remove from I any a satisfying that $\rho(a, I \setminus \{a\}) \geq \phi$. In the end, $|I| \geq \left(1 - \frac{\varphi}{2\phi}\right)|A|$.

2.2 Greedy Inductively Independent Set

In this subsection, we present two lower bounds on the size of the φ-restricted greedy IIS J of A in an ordering $\prec$ of A for any $\varphi > 0$. Let $\succ$ denote the reverse of the ordering $\prec$.

Theorem 2. *Suppose that A is a ϕ-restricted IIS in $\succ$ for some $\phi > 0$. Then,*

$$|J| \geq \frac{|A|}{1 + \phi/\varphi}.$$

Proof. By the greedy principle, for each $b \in A \setminus J$,

$$\rho(J_{\prec b}, b) \geq \varphi.$$

On the other hand, for any $a \in A$,

$$\rho(A_{\succ a}, a) < \phi.$$

Thus,

$$\begin{aligned}\varphi|A \setminus J| &\leq \sum_{b \in A \setminus J} \rho(J_{\prec b}, b) = \sum_{a \in J} \rho\left(a, (A \setminus J)_{\succ a}\right) \\ &\leq \sum_{a \in J} \rho(a, A_{\succ a}) < |J|\phi,\end{aligned}$$

which implies

$$|A \setminus J| < \frac{\varphi}{\phi}|J|.$$

So,

$$|A| = |J| + |A \setminus J| < (1 + \phi/\varphi)|J|,$$

and hence

$$|J| > \frac{|A|}{1 + \phi/\varphi}.$$

Thus, the theorem holds.

Consider a positive parameter γ. The *forward γ-shield number* of a set $B \subseteq A$ in $\prec$ is defined to be the smallest integer k satisfying that for any $B' \subseteq B$ and any $a \in A$ with $a \prec B'$, there exists $S \subseteq B'$ satisfying that $|S| \leq k$ and for each $b \in B' \setminus S$,

$$\rho(a, b) \leq \gamma\rho(S, b).$$

Theorem 3. *Suppose that O is a ϕ-restricted IIS of A in $\prec$ for some $\phi > 0$ and its forward $\frac{\varphi}{\phi}$-shield number in $\prec$ is μ. Then, $|J| \geq |O| / \mu$.*

Proof. We prove the theorem by contradiction. Assume to the contrary that $|J| < |O| / \mu$. We construct a partition of O as follows. First, $O \cap J$ is partitioned into singleton subsets, and let $S(a) = \{a\}$ for each $a \in O \cap J$. Then, $O \cap J$ is partitioned as follows. Initialize O' to $O \setminus J$. For each $a \in J \setminus O$ in the order of addition to J, let $S(a)$ be a minimum subset $S \subseteq O'_{\succ a}$ satisfying that each $b \in O'_{\succ a} \setminus S$,

$$\rho(a,b) \leq \frac{\varphi}{\phi} \rho(S,b);$$

and remove $S(a)$ from O'. As $|S(a)| \leq \mu$ for each $a \in J \setminus O$ and $|O| > \mu |J|$, the final O' is non-empty. After this partition, we choose an arbitrary link $b \in O'$. Then, $J_{\prec b} \neq \emptyset$, and all of $S(a)$ for $a \in J_{\prec b}$ are non-empty as well. However,

$$\begin{aligned}\rho(J_{\prec b}, b) &= \textstyle\sum_{a \in J_{\prec b}} \rho(a,b) \leq \frac{\varphi}{\phi} \textstyle\sum_{a \in J_{\prec b}} \rho(S(a), b) \\ &= \frac{\varphi}{\phi} \rho\left(\textstyle\bigcup_{a \in J_{\prec b}} S(a), b\right) < \frac{\varphi}{\phi} \phi = \varphi.\end{aligned}$$

Hence, b should have been added to J by the greedy principle, which is a contradiction.

3 Relative Inteference System

Suppose that A is a non-empty subset A of L and p is a power assignment on A satisfying that $p_0(a) < p(a) \leq P$. The relative intereference between a and b under p is defined to be

$$\overline{RI}_p(a,b) = RI_p(a,b) + RI_p(b,a).$$

The symmetric conflict system $\left(A, \overline{RI}_p\right)$ is referred to as the relative interference system of (A,p). For any $\phi > 0$, and a ϕ-restricted IS of A w.r.t. $\overline{RI}_p$ is also referred to as a ϕ-restricted *strongly independent set* (SIS) of A under p. In this section, we explore the structual properties of $\left(A, \overline{RI}_p\right)$.

Consider a positive parameter β. For a link $a \in A$ and a set B of disjoint links in A, a subset S of B is said to be a *β-guard* of B from a if for each link $b \in B \setminus S$,

$$\begin{aligned}\min_{a' \in S} \ell(b,a') &\leq \beta \ell(b,a), \\ \min_{a' \in S} \ell(a',b) &\leq \beta \ell(a,b).\end{aligned}$$

The *β-guard number* of a set $B \subseteq A$ is defined to be the smallest integer k satisfying that for any $B' \subseteq B$ and any $a \in A$, there is a β-guard of B' from a whose size is at most k.

Let I be a ϕ-restricted SIS of A under p for some $\phi > 0$. Consider an integer $m > \phi/\sigma$ and let

$$\beta = \frac{2}{1 - (m\sigma/\phi)^{-1/\kappa}}.$$

For any $a \in I$, we construct a β-guard S of I from a as follows. Choose b_1 to be a link in I whose receiver is closest to the receiver of a, and b_2 to be a link in I whose sender is closest to the sender of a. Let

$$\begin{aligned} S_1 &= \{b \in I \setminus \{b_1\} : \ell(b, b_1) > \beta\ell(b, a)\}, \\ S_2 &= \{b \in I \setminus \{b_2\} : \ell(b_2, b) > \beta\ell(a, b)\}, \\ S &= \{b_1\} \cup \{b_2\} \cup S_1 \cup S_2. \end{aligned}$$

Clearly, S is a β-guard of I from a. The next theorem presents an upper bound on $|S|$.

Theorem 4. *Both $|S_1|$ and $|S_2|$ are at most m, and hence $|S|$ is at most $2(m+1)$.*

Proof. Let $a = (u_0, v_0)$ and $b_i = (u_i, v_i)$ for $i = 1, 2$.

We first prove $|S_1| \leq m$ by contradiction. Assume to the contrary that $|S_1| \geq m+1$. Then, $v_1 \neq v_0$, for otherwise for any link $b \in I \setminus \{b_1\}$, $\ell(b, b_1) = \ell(b, a)$ and hence S_1 is empty. We then observe that the sender of each link in S_1 is apart from v_0 by less than $\|v_0v_1\| / (\beta - 1)$. Indeed, for each link $b \in S_1$,

$$\|v_0v_1\| \geq \ell(b, b_1) - \ell(b, a) > (\beta - 1)\ell(b, a)$$

and hence

$$\ell(b, a) < \|v_0v_1\| / (\beta - 1).$$

Let $b' = (u', v')$ be a link in S_1 with the smallest *transmission* power $p(b')$. Then,

$$\begin{aligned} \|u'v'\| &\geq \|v_0v'\| - \|v_0u'\| \\ &> \|v_0v_1\| - \frac{1}{\beta - 1}\|v_0v_1\| \\ &= \frac{\beta - 2}{\beta - 1}\|v_0v_1\|. \end{aligned}$$

Thus, for any $b = (u, v) \in S_1 \setminus \{b'\}$,

$$\begin{aligned} \frac{\ell(b')}{\ell(b, b')} &= \frac{\|u'v'\|}{\|uv'\|} \geq \frac{\|u'v'\|}{\|v_0u\| + \|v_0u'\| + \|u'v'\|} \\ &\geq \frac{\|u'v'\|}{\frac{2}{\beta-1}\|v_0v_1\| + \|u'v'\|} = \frac{1}{\frac{2}{\beta-1}\frac{\|v_0v_1\|}{\|u'v'\|} + 1} \\ &\geq \frac{1}{\frac{2}{\beta-1}\frac{1}{\frac{\beta-2}{\beta-1}} + 1} = \frac{\beta - 2}{\beta} = (m\sigma/\phi)^{-1/\kappa}, \end{aligned}$$

and consequently

$$RI_p\left(b,b'\right) > \sigma \frac{p\left(b\right)}{p\left(b'\right)} \left(\frac{\ell\left(b'\right)}{\ell\left(b,b'\right)} \right)^{\kappa} \geq \sigma \left(\left(m\sigma/\phi\right)^{-1/\kappa} \right)^{\kappa} = \frac{\phi}{m}.$$

So,

$$\phi > RI_p\left(S_1 \setminus \left\{b'\right\}, b'\right) \geq \left(|S_1| - 1\right) \frac{\phi}{m} \geq m \frac{\phi}{m} = \phi,$$

which is a contradiction. Therefore, $|S_1| \leq m$.

Next, we prove $|S_2| \leq m$ by contradiction. Assume to the contrary that $|S_2| \geq m+1$. Then, $u_2 \neq u_0$, for otherwise for any link $b \in I \setminus \{b_2\}$, $\ell\left(b_2, b\right) = \ell\left(a, b\right)$ and hence S_2 is empty. We then observe that the receiver of each link in S_2 is apart from u_0 by less than $\left\|u_0 u_2\right\| / \left(\beta - 1\right)$. Indeed, for each link $b \in S_2$,

$$\left\|u_0 u_2\right\| \geq \ell\left(b_2, b\right) - \ell\left(a, b\right) > \left(\beta - 1\right) \ell\left(a, b\right),$$

and hence

$$\ell\left(a, b\right) < \left\|u_0 u_2\right\| / \left(\beta - 1\right).$$

Let $b' = \left(u', v'\right)$ be a link in S with the largest *reception* power $p\left(b'\right) \ell\left(b'\right)^{-\kappa}$. Then,

$$\left\|u'v'\right\| \geq \left\|u_0 u'\right\| - \left\|u_0 v'\right\| > \left\|u_0 u_2\right\| - \frac{1}{\beta - 1} \left\|u_0 u_2\right\| = \frac{\beta - 2}{\beta - 1} \left\|u_0 u_2\right\|.$$

Thus, for any $b = \left(u, v\right) \in S_2 \setminus \{b'\}$,

$$\begin{aligned}
\frac{\ell\left(b'\right)}{\ell\left(b', b\right)} &= \frac{\left\|u'v'\right\|}{\left\|u'v\right\|} \geq \frac{\left\|u'v'\right\|}{\left\|u_0 v\right\| + \left\|u_0 v'\right\| + \left\|u'v'\right\|} \\
&\geq \frac{\left\|u'v'\right\|}{\frac{2}{\beta-1} \left\|u_0 u_2\right\| + \left\|u'v'\right\|} = \frac{1}{\frac{2}{\beta-1} \frac{\left\|u_0 u_2\right\|}{\left\|u'v'\right\|} + 1} \\
&\geq \frac{1}{\frac{2}{\beta-1} \frac{1}{\frac{\beta-2}{\beta-1}} + 1} = \frac{\beta - 2}{\beta} = \left(m\sigma/\phi\right)^{-1/\kappa},
\end{aligned}$$

and consequently

$$\begin{aligned}
RI_p\left(b', b\right) &> \sigma \frac{p\left(b'\right)}{p\left(b\right)} \left(\frac{\ell\left(b\right)}{\ell\left(b', b\right)} \right)^{\kappa} = \sigma \frac{p\left(b'\right) \ell\left(b'\right)^{-\kappa}}{p\left(b\right) \ell\left(b\right)^{-\kappa}} \left(\frac{\ell\left(b'\right)}{\ell\left(b', b\right)} \right)^{\kappa} \\
&\geq \sigma \left(\frac{\ell\left(b'\right)}{\ell\left(b', b\right)} \right)^{\kappa} \geq \sigma \left(\left(m\sigma/\phi\right)^{-1/\kappa} \right)^{\kappa} = \frac{\phi}{m}.
\end{aligned}$$

So,

$$\phi > RI_p\left(S_2 \setminus \left\{b'\right\}, b'\right) \geq \left(|S_2| - 1\right) \frac{\phi}{m} \geq m \frac{\phi}{m} = \phi,$$

which is a contradiction. Therefore, $|S_2| \leq m$.

Since any subset of a ϕ-restricted SIS is also a ϕ-restricted SIS, we have the following theorem.

Theorem 5. *The β-guard number of any ϕ-restricted SIS of A under p is at most* $2(m+1)$

By choosing $m = 1$ in Theorem 5, we immediately get the following corollary.

Corollary 1. *Consider a parameter $\phi \in (0, \sigma)$. Let*

$$\beta = \frac{2}{1 - (\sigma/\phi)^{-1/\kappa}}.$$

Then, the β-guard number of any ϕ-restricted SIS of A under p is at most 4.

By choosing $\phi = \frac{1}{2\left(2+(2\sigma)^{-1/\kappa}\right)^{\kappa}}$ in the above corollary, we further have the following corollary.

Corollary 2. *For $\phi = \frac{1}{2\left(2+(2\sigma)^{-1/\kappa}\right)^{\kappa}}$, the $(2\phi)^{-1/\kappa}$-guard number of of any ϕ-restricted SIS of A under p is at most* 4.

4 MISL with Monotone and Sublinear Power Assignment

Consider a non-empty subset A of L and a monotone and sublinear power assignment p on A satisfying that for each $a \in A$, $p_0(a) < p(a) \leq P$. Let $\prec$ denote an ordering of A in the the *increasing* order of length and $\succ$ denote the reverse of $\prec$. We consider the conflict system $\left(A, \overline{RI}_p\right)$ and present a relaxation-based algorithm **RelaxIS** for computing a 1-restricted SIS of A under p. The algorithm proceeds in two steps:

- **Step 1**: Compute the greedy $\frac{1}{2}$-restricted IS J of A in $\prec$ with respect to $\overline{RI}_p$.
- **Step 2**: Compute the greedy $\frac{1}{2}$-restricted IS I of J in $\succ$ with respect to $\overline{RI}_p$.

Clearly, the set I is a 1-restricted SIS of A under p. The theorem below gives an approximation bound of this algorithm.

Theorem 6. *The algorithm **RelaxIS** has an approximation bound* $64\left(2 + (2\sigma)^{-1/\kappa}\right)^{\kappa}$.

The proof of the above theorem utilizes the following property of the montone and sublinear power assignment p.

Theorem 7. *Suppose that B is a set of disjoint links in A. Then, for any $\beta > 0$ the forward β^{κ}-shield number of B in $\prec$ w.r.t. $\overline{RI}_p$ is no more than the β-guard number of B.*

Proof. Let μ be the β-guard number of B. Consider any $B' \subseteq B$ and any $a \in A$ with $a \prec B'$. Then, there is a β-guard S of B' from a whose size is at most μ.

Consider any $b \in B' \setminus S$. There exist two (possibly identical) links b_1 and b_2 in S satisfying that

$$\begin{aligned}\ell(b, b_1) &\leq \beta \ell(b, a),\\ \ell(b_2, b) &\leq \beta \ell(a, b).\end{aligned}$$

As all links in B are disjoint, both $\ell(b, b_1)$ and $\ell(b_2, b)$ are positive, so are $\ell(b, a)$ and $\ell(a, b)$. Since p is sublinear and $a \prec b_1$, we have

$$\begin{aligned}\frac{RI_p(b,a)}{RI_p(b,b_1)} &= \left(\frac{\ell(b,b_1)}{\ell(b,a)}\right)^{\kappa} \frac{(p(b_1) - p_0(b_1))\,\ell(b_1)^{-\kappa}}{(p(a) - p_0(a))\,\ell(a)^{-\kappa}}\\ &= \left(\frac{\ell(b,b_1)}{\ell(b,a)}\right)^{\kappa} \frac{\left(p(b_1)\,\ell(b_1)^{-\kappa} - \sigma\xi_0\right)}{\left(p(a)\,\ell(a)^{-\kappa} - \sigma\xi_0\right)}\\ &\leq \left(\frac{\ell(b,b_1)}{\ell(b,a)}\right)^{\kappa} \leq \beta^{\kappa}.\end{aligned}$$

Since p is monotone and $a \prec b_2$, we have

$$\frac{RI_p(a,b)}{RI_p(b_2,b)} = \left(\frac{\ell(b_2,b)}{\ell(a,b)}\right)^{\kappa} \frac{p(a)}{p(b_2)} \leq \left(\frac{\ell(b_2,b)}{\ell(a,b)}\right)^{\kappa} \leq \beta^{\kappa}.$$

Thus,

$$\begin{aligned}\overline{RI}_p(a,b) &\leq \beta^{\kappa}\left(RI_p(b,b_1) + RI_p(b_2,b)\right)\\ &\leq \beta^{\kappa}\,\overline{RI}_p(\{b_1\} \cup \{b_2\}, b)\\ &\leq \beta^{\kappa}\,\overline{RI}_p(S, b).\end{aligned}$$

Since $|S| \leq \mu$, the theorem holds.

Theorem 7 together with Corollary 2 implies the following corollary.

Corollary 3. *For $\phi = \frac{1}{2\left(2+(2\sigma)^{-1/\kappa}\right)^{\kappa}}$, the forward $\frac{1}{2\phi}$-shield number of any ϕ-restricted SIS of A under p in $\prec$ is at most* 4.

Now, we are ready to prove Theorem 6. Consider a maximum IS O of A under p. Let

$$\phi = \frac{1}{2\left(2 + (2\sigma)^{-1/\kappa}\right)^{\kappa}}$$

and O' be a maximum ϕ-restricted SIS of O under p. By Theorem 1, $|O'| \geq \frac{|O|}{4\phi}$. By Corollary 3, the forward $\frac{1}{2\phi}$-shield number of O' in $\prec$ is at most 4. So, by Theorem 3, $|J| \geq \frac{|O'|}{4}$. By Theorem 2,

$$|I| \geq \frac{|J|}{2} \geq \frac{|O'|}{8} \geq \frac{|O|}{32\phi} = \frac{|O|}{64\left(2 + (2\sigma)^{-1/\kappa}\right)^{\kappa}}.$$

So, Theorem 6 holds.

Finally, we remark that **Step 2** of the algorithm **RelaxIS** can be replaced by applying the algorithm **ExtractIS**(1) to extract a 1-restricted SIS I from J. With such modification, Theorem 6 still holds.

Acknowledgements. This work was supported by the National Plan for Sciences & Technology (NPST) at King Saud University (Project No. 10-INF1184-02), the National Science Foundation of USA under grants CNS-0916666 and CNS-1219109, and by the National Natural Science Foundation of P. R. China under grant 61128005.

References

1. Andrews, M., Dinitz, M.: Maximizing Capacity in Arbitrary Wireless Networks in the SINR Model: Complexity and Game Theory. In: IEEE INFOCOM (2009)
2. Chafekar, D., Kumar, V., Marathe, M., Parthasarathy, S., Srinivasan, A.: Approximation algorithms for computing capacity of wireless networks with SINR constraints. In: IEEE INFOCOM, pp. 1166–1174 (2008)
3. Goussevskaia, O., Oswald, Y.A., Wattenhofer, R.: Complexity in geometric SINR. In: Proc. of the 8th ACM MOBIHOC, pp. 100–109 (September 2007)
4. Halldórsson, M.M., Mitra, P.: Wireless capacity with oblivious power in general metrics. In: SIAM SODA, pp. 1538–1548 (2011)
5. Kesselheim, T.: A Constant-Factor Approximation for Wireless Capacity Maximization with Power Control in the SINR Model. In: SIAM SODA, pp. 1549–1559 (2011)
6. Wan, P.-J., Jia, X., Yao, F.: Maximum Independent Set of Links under Physical Interference Model. In: Liu, B., Bestavros, A., Du, D.-Z., Wang, J. (eds.) WASA 2009. LNCS, vol. 5682, pp. 169–178. Springer, Heidelberg (2009)
7. Wan, P.-J., Ma, C., Tang, S., Xu, B.: Maximizing Capacity with Power Control under Physical Interference Model in Simplex Mode. In: Cheng, Y., Do Eun, Y., Qin, Z., Song, M., Xing, K. (eds.) WASA 2011. LNCS, vol. 6843, pp. 84–95. Springer, Heidelberg (2011)

An Auction Mechanism for Resource Allocation in Mobile Cloud Computing Systems

Yang Zhang, Dusit Niyato, and Ping Wang

School of Computer Engineering, Nanyang Technological University (NTU),
Singapore

Abstract. A mobile cloud computing system is composed of heterogeneous services and resources to be allocated by the cloud service provider to mobile cloud users. On one hand, some of these resources are substitutable (e.g., users can use storage from different places) that they have similar functions to the users. On the other hand, some resources are complementary that the user will need them as a bundle (e.g., users need both wireless connection and storage for online photo posting). In this paper, we first model the resource allocation process of a mobile cloud computing system as an auction mechanism with premium and discount factors. The premium and discount factors indicate complementary and substitutable relations among cloud resources provided by the service provider. Then, we analyze the individual rationality and incentive compatibility (truthfulness) properties of the users in the proposed auction mechanism. The optimal solutions of the resource allocation and cost charging schemes in the auction mechanism is discussed afterwards.

Keywords: Mobile cloud computing, auction, mechanism design.

1 Introduction

In a cloud computing system, the service provider has different resources to be leased or sold to cloud users. The users use the allocated resources to run their applications. A mobile cloud computing system differs from general cloud computing systems in some aspects. One important aspect is the combination pattern of demanded services. In a general cloud application, a cloud user may request either a single service or a combination of services. For example, a user of Amazon's EC2 [1] may subscribe for a server only, or spend extra cost to subscribe for other resources and services only when necessary. However, services in a mobile cloud are generally provided in bundles. The reason is that a mobile device as an end user is relatively a "thin client" which cannot process too complex tasks. Therefore, most of the tasks are offloaded to the cloud side, and the cloud service providers need to provide a bundle of service to process the tasks. For example, in a mobile game service, the user will need both a processing resource for artificial intelligent player and a storage for game module. Most importantly, communication bandwidth, as crucial resource in mobile systems, should be guaranteed for transferring data of mobile cloud applications.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 76–87, 2013.

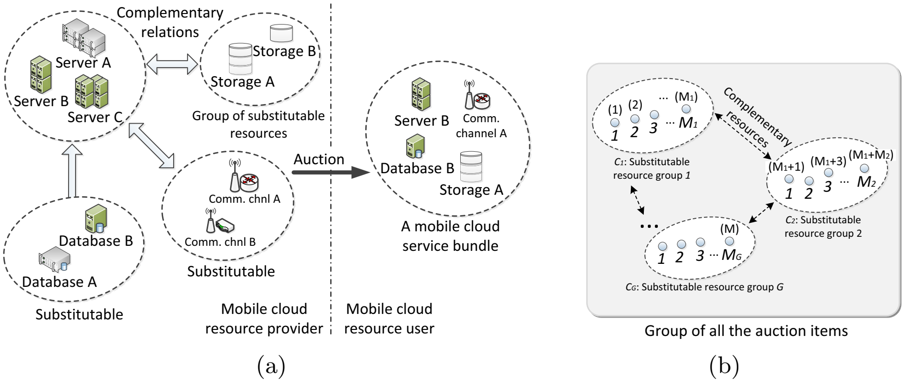


Fig. 1. Asymmetric scenario: (a) user 1's utility to her type t_1, (b) user 2's utility to user 1's type t_1, and (c) the service provider's revenue to user 1's type t_1. Premium and discount factors' impacts: (d) user's utility under different settings of premium and discount factors.

In this paper, we model the resource allocation problem in a mobile cloud system as a combinatorial auction with substitutable or complementary commodities. The service provider is defined as a seller, while the users are defined as buyers. The available cloud resources are to be sold and allocated by the seller to buyers. Moreover, we design the proposed auction mechanism for the mobile cloud computing system, which is proved to be individual rational and incentive compatible (as shown in Section 4.1). As shown in Fig. 1(a), the cloud resources of the service provider can be categorized into several groups, e.g., processing (i.e., server), storage, and communication resources. Resources in the same group are different in quality, but have similar functions. These heterogeneous resources can be allocated in bundles. Therefore, we consider the resources as commodities which could be substitutable or complementary. If the resources are substitutable, the valuations of these resources are sub-additive (i.e., total valuation of all resources obtained altogether is less than the sum of valuation of each resource). For example, any server can be treated by the user as a substitutable resource. On the other hand, if the resources are complementary (e.g., a server for processing and bandwidth for data transmission where the user needs both), the valuations of these resources are super-additive (i.e., total valuation of resources obtained altogether is higher than the sum of valuation of each resource).

The rest of this paper is organized as follows. In Section 2, we review related work. We describe the system model in Section 3. In Section 4, we discuss the individual rationality and the incentive compatibility properties of cloud users. Also, the optimal solutions of the service provider's resource allocation and cost charging schemes are presented. Section 5 shows the numerical results. Finally, we conclude and summarize the paper in Section 6.

2 Related Work

In mobile cloud computing, the tasks of mobile applications will be partly offloaded to servers in cloud. Therefore, the performance of mobile applications can be improved (e.g. computer games [2]). A few works addressed the issues of general/mobile cloud computing systems. In [3], the definition and discussion of general cloud computing systems were presented. [4] focused on cloud computing implemented on mobile network infrastructures, i.e., mobile cloud computing systems. Also, [5] presented a comprehensive survey of mobile cloud computing including the system architecture, applications, resource allocation and other issues.

Auction is a general and effective way for resource allocation. A tutorial introduction of auction theory for computer science was presented in [10]. Specifically, [6] developed an auction mechanism for a single auction commodity scenario, and proposed the concepts of individual rationality and incentive compatibility in auction mechanism designs. [7] extended the model in [6] to a multiple commodity auction scenario. However the auction commodities in [7] are independent with each other and indivisible. [8] discussed an auction model with two complementary auction commodities. [9] explored the auction designs for substitutable commodities. [11] proposed an auction technique to optimize cloud resource distribution in a cloud computing system. [12] proposed a second-price auction mechanism to allocate a single type of cloud resource, i.e., computational capacity.

To our best knowledge, the analytical auction model for a mobile cloud computing system containing both complementary and substitutable resource was not considered before.

3 System Model and Assumptions

In this section, we first present the description of the mobile cloud computing system and auction mechanism. Then, the utility of users is defined.

3.1 Mobile Cloud Computing System and Auction Mechanism

We consider a mobile cloud computing system with a service provider offering mobile cloud applications/services to N users (Fig. 1(a)). The service provider has cloud resources (e.g., CPU/computational capacity, database, and communications) to support the offered services. There are totally M resources which are divided into G groups. Group k contains M_k resources, for $k = 1, 2, \ldots, G$. In the same group, the resources provide similar functions, and hence the resources are substitutable. On the other hand, in different groups, the resources provide different functions, and hence the resources are complementary in building cloud services for the users.

The substitutable and complementary resources in a mobile cloud computing system is shown in Fig. 1(b) which is an abstracted model of Fig. 1(a).

We assume that the transitivity condition of the substitutable and complementary resources holds, i.e.,

- Resources A, B and C are substitutable/complementary mutually, if and only if, resources A and B, as well as B and C are substitutable/complementary.
- Resources A and C are complementary, if resources A and B are complementary, while resources B and C are substitutable.

By assuming the transitivity condition, there is no "triangle" relations among any three resources A, B and C such that A and B are substitutable (in one group), B and C are complementary (in different groups), but A and C are still substitutable.

The auction mechanism is developed for resource allocation by the mobile cloud service provider to the resource buyers (i.e., users). The auction mechanism works as follows. First, the users submit bids containing users' valuations on the resources to the service provider. The valuation of user i on resource j is denoted by v_{ij}. There are other users who also compete for the resources. Therefore, the valuation depends not only on the type (i.e., preference) of user i, but also on other users' types. The user i's type, denoted by t_i, is defined as the user's private appetite on obtaining the resources. Therefore, the valuation is defined as $v_{ij}(t_i)$, where $\mathbf{t}_{-i}$ is the vector of other users' types except that of user i and $\mathbf{t} = (t_i) = (t_1, t_2, \ldots, t_N)$. The service provider receives the valuation but does not know the user's real type. After receiving the valuations for the resources from all users, the service provider then optimizes the resource allocation to maximize the revenue. The service provider sends back the allocation result denoted by (p, c), where p is the set of all allocation p_{ij}, and $c = \{c_1, c_2, \ldots, c_N\}$ is the set of the costs charged to users. p_{ij} is the proportion of resource j allocated to user i. $c_i(t_i)$ denotes the cost charged by the service provider to user i. Again, the cost is ultimately a function of user's type. Formally, we define (p, c) as a *mechanism* revealed by the service provider to all users.

3.2 User's Utility

A user has a satisfaction on the allocated resources, which is referred to as the utility. The expected utility of user i to obtain the resources from a group k of substitutable resources can be defined as follows:

$$u_i^{k,s}(p_{i\cdot}^k, c_i^k, t_i) = \mathbb{E}_{\mathbf{t}_{-i}}\Big\{ \sum_{r=1}^{M_k} p_{ir}^k(t_i) v_{ir}^k(t_i) - l_i(t_i) \sum_{r=1}^{M_k} p_{ir}^k(t_i) \Big\} - c_i^k(t_i) \qquad (1)$$

where t_i is the type of user i, and $t_i \in [\underline{t_i}, \overline{t_i}]]$. $\underline{t_i}$ and $\overline{t_i}$ are the lower bound and upper bound of t_i, respectively. In this case, we assume that the type of a user is a random variable with cumulative distribution function (CDF) denoted by $F(t_i)$ and probability density function (PDF) denoted by $f(t_i)$. CDF and PDF are known knowledge or observations to the service provider. p_{ir}^k denotes the proportion of rth resource in group k allocated to user i. $p_{i\cdot}^k$ is p_{ir}^k for all r. v_{ir}^k is user i's private valuation function on the rth resource of group k. We assume that the valuation functions are positive, non-decreasing and concave with respect to t_i. c_i^k is the total cost function of all the resources in group k for user i. This cost is paid by user i to the service provider. Note that $l_i(t_i)\sum_{r=1}^{M_k} p_{ir}^k(t_i)$ in (1) represents the sub-additive term of the substitutable resources. The sub-additive term indicates the loss of utility of a user when multiple resources are allocated from the same group of substitutable resources. Here $l_i(t_i)$ is defined as a discount factor. For example, if the discount factor is large, the user has more dissatisfaction on obtaining the resources from the same group.

Next, we consider the complementary resources and their contributions in the utility function. First, we define the amount of aggregated resources allocated from group k from the service provider as follows:

$$p^{k,c} = \sum_{r=1}^{M_k} p_{ir}^k(t_i). \tag{2}$$

Then, the total utility of user i from obtaining resources can be expressed as follows:

$$u_i(t_i) = \mathbb{E}_{\mathbf{t}_{-i}}\Big\{\sum_{k=1}^{G} u_i^{k,s}(p_{i\cdot}^k, c_i^k, t_i) + h_i(t_i)\prod_{k=1}^{G} p^{k,c}(t_i)\Big\} \tag{3}$$

where $h_i(t_i)\prod_{k=1}^{G} p^{k,c}(t_i)$ defined in (3) represents the super-additive term of the complementary resources. $h_i(t_i)$ is defined as a premium factor. If the premium factor is large, the user has more satisfaction on obtaining the resources from the different groups. Note that if the user has received nothing from merely one group, the super-additive term will become zero, as a punishment of not receiving a complete bundle of complementary resources required for a mobile cloud computing service.

We will derive the total utility of all M resources next. We substitute (1) and (2) into (3). In this case, the variables with k are mapped to the new variables without k. The process is shown in Fig. 1(b). That is, the rth resource in group k is mapped to the jth resource regardless of the group, where $j = \sum_{i=1}^{k-1} M_k + r$. After mapping, p_{ir}^k, v_{ir}^k and c_i^k become p_{ij}, v_{ij} and c_i, respectively. p_{ij}, v_{ij} and c_i have the same definitions as those in Section 3.1. Then we can express the utility $u_i(p_{i\cdot}, c_i, t_i)$ of the user i as follows:[1]

$$u_i(t_i) = \mathbb{E}_{\mathbf{t}_{-i}}\Big\{\sum_{j=1}^{M} p_{ij}(t_i)v_{ij}(t_i) + S_i(t_i)\Big\} - c_i(t_i) \tag{4}$$

where $S_i(t_i, \mathbf{t}_{-i})$ includes the premium and discount terms defined as follows:

$$S_i(t_i) = h_i(t_i)\prod_{k=1}^{G}\sum_{j=\sum_{n=1}^{k-1} M_k+1}^{\sum_{n=1}^{k} M_n} p_{ij}(t_i) - l_i(t_i)\sum_{j=1}^{M} p_{ij}(t_i). \tag{5}$$

[1] The notation $u_i(t_i)$ is used for short.

We assume in our model that the valuation and extra premium and discount information $S_i(t_i, \mathbf{t}_{-i})$ are sent to the service provider. Therefore, the first term in (4) could be treated as user i's "virtual valuation" with premium and discount factors. Note, however, that the valuations submitted by users to the service provider may not be the true valuations on the resources. For example, to achieve a higher profit, user i who has the valuation $v_{ij}(t_i)$ on resource j might actually submit a falsified valuation $v_{ij}(\hat{t}_i)$ to the service provider as if the user had the fake type $\hat{t}_i$. Only in the auction mechanism (p, c) that guarantees truthfulness, the users will choose to submit their true valuations as the best bidding strategies. By sending an untruthful valuation, the user i's utility is

$$\hat{u}_i(\hat{t}_i|t_i) = \mathbb{E}_{\mathbf{t}_{-i}}\Big\{\sum_{j=1}^{M} p_{ij}(\hat{t}_i)v_{ij}(t_i) + S_i(\hat{t}_i)\Big\} - c_i(\hat{t}_i). \tag{6}$$

4 Analysis of the Auction Mechanism for Mobile Cloud Computing System

In this section, the individual rationality and the incentive compatibility properties of the auction mechanism are analyzed. The optimization of the revenue of the service provider is proposed. Also, the structure of the utility function is discussed.

4.1 Individual Rationality and Incentive Compatibility

When an auction and resource allocation mechanism is designed, the auctioneer must ensure positive payoffs of the auction participants (i.e., buyers), so that those buyers are willing to join the auction market. On the other hand, the mechanism should discourage the buyers to submit valuations which are not based on the buyers' true valuation. The individual rationality and incentive compatibility (truthfulness) properties are defined as follows:

$$\begin{aligned} \text{Individual rationality :} \quad & u_i(t_i) \geq 0 \\ \text{Incentive compatibility :} \quad & u_i(t_i) \geq \hat{u}_i(\hat{t}_i|t_i). \end{aligned} \tag{7}$$

Proposition 1. *In the proposed auction for the substitutable and complementary resources in a mobile cloud computing system, a mechanism* (p, c) *is individually rational and incentive compatible if and only if the following conditions hold,*

$$u_i(\underline{t_i}) \geq 0 \tag{8}$$

$$u_i(t_i) = u_i(\underline{t_i}) + \mathbb{E}_{\mathbf{t}_{-i}}\Big\{\sum_{j=1}^{M}\int_{\underline{t_i}}^{t_i} \frac{\partial v_{ij}(x)}{\partial t_i} p_{ij}(x)dx\Big\} \tag{9}$$

$$\mathbb{E}_{\mathbf{t}_{-i}}\Big\{\sum_{j=1}^{M}\int_{\hat{t}_i}^{t_i} \frac{\partial v_{ij}(x)}{\partial t_i} p_{ij}(x)dx\Big\} \geq \mathbb{E}_{\mathbf{t}_{-i}}\Big\{\sum_{j=1}^{M}\big(v_{ij}(t_i) - v_{ij}(\hat{t}_i)\big)p_{ij}(\hat{t}_i)\Big\}. \tag{10}$$

The proof of Proposition 1 could be done by expand the individual rationality and the incentive compatibility expressions with the definition of (real/fake) utility 4 substituted. The detailed proof procedure is provided in the extended paper [13].

4.2 The Seller-Side Problems: Revenue Maximization and Cost Charging

The objective of the service provider is to sell and allocate available resources to the users such that the revenue is maximized. The revenue of the service provider is the sum of the cost $c_i(t_i)$ that each user pays for the allocated resources. The mechanism (i.e., resource allocation) is designed to maximize the total revenue of the service provider, i.e. $\sum_{i=1}^{N} \mathbb{E}_{t_i}\{c_i(t_i)\}$. The following proposition states the optimal mechanism (i.e., maximizing the revenue of the service provider). The following proposition is proved in [13].

Proposition 2. *(p^*, c^*) is an optimal mechanism if p^* (allocation scheme) maximizes*

$$\sum_{i=1}^{N} \mathbb{E}_{\mathbf{t}}\Big\{ \sum_{j=1}^{M} \big(v_{ij}(t_i) - \frac{1-F(t_i)}{f(t_i)} \frac{\partial v_{ij}(t_i)}{\partial t_i}\big) p_{ij}(t_i) + S_i(t_i) \Big\} \tag{11}$$

subject to the constraints in (8), (9) and (10).

To achieve the optimal revenue, the (maximum) cost, i.e., payment, $c_i^*(t_i)$ charged to each user i can be obtained from

$$c_i^*(t_i) = \mathbb{E}_{\mathbf{t}}\Big\{ \sum_{j=1}^{M} p_{ij}^*(t_i) v_{ij}(t_i) - \sum_{j=1}^{M} \int_{\underline{t_i}}^{t_i} \frac{\partial v_{ij}(x)}{\partial t_i} p_{ij}^*(x) dx + S_i^*(t_i) \Big\}. \tag{12}$$

Note that the cost charging and revenue optimization as in Proposition 2 are essentially linear programming problems, which can be solved numerically by the service provider (seller), provided that the seller has the knowledge of CDF and PDF of buyers' types as we assumed in Section 3.2.

4.3 Structures of Utility and Revenue Functions

We analyze the structures of the users' utility and the service provider's optimal revenue functions. According to (9), each user's utility is only decided by the "basic utility" (i.e., $u_i(\underline{t_i})$), and the "marginal utility" (i.e., $\mathbb{E}_{\mathbf{t}_{-i}}\big\{ \sum_{j=1}^{M} \int_{\underline{t_i}}^{t_i} \frac{\partial v_{ij}(x)}{\partial t_i} p_{ij}(x) dx \big\}$).

In the optimization process of the provider's total revenue, as justified in the proof of Proposition 2 ([13]), the basic utility is minimized to be zero by the service provider's resource allocation. Therefore, each user's utility depends solely on the marginal utility.

From another point of view, from the expression of $\sum_{i=1}^{N} \mathbb{E}_{t_i}\{u_i(t_i)\}$ in the proof of Proposition 2 ([13]), the user's optimal utility after the allocation can be expressed as follows:

$$u_i^*(t_i) = \mathbb{E}_{\mathbf{t}}\Big\{ \sum_{j=1}^{M} \frac{1 - F(t_i)}{f(t_i)} \frac{\partial v_{ij}(t_i)}{\partial t_i} p_{ij}^*(t_i) \Big\}. \tag{13}$$

As we have assumed, the valuation functions of users are non-decreasing and concave. Therefore, $\frac{\partial v_{ij}(t_i)}{\partial t_i}$ decreases with respect to t_i. Then, the utility is affected by $\frac{1-F(t_i)}{f(t_i)}$ which is the distribution pattern of user's type and $p_{ij}^*(t_i)$ which is the amount of resources allocated to the user.

For the service provider, the optimal revenue from selling resources can be calculated by placing the optimal allocation p^* into the revenue function defined in (11). According to the second term in (11), the service provider's revenue is also affected by the marginal utility of the users. Moreover, the premium and discount term (i.e., $S_i(t_i)$) affects the optimal revenue, depending on users' choices of coefficients h_i and l_i (see (5)).

In the following section, we will use examples and numerical results to support the analyses of utility and revenue functions.

5 Example Scenarios and Numerical Results

In this section, we present examples and numerical simulation results of the proposed auction model for the substitutable and complementary resources. We discuss a model with two mobile cloud users for simplicity. The mobile cloud service provides two servers for computational tasks, as well as two communication channels to support different transmission rate (or bandwidth) requirements. We consider a symmetric and an asymmetric scenarios. Then, we show the impact caused by the premium and discount factors to the user's utility and service provider's revenue.

Suppose there are four commodities to be auctioned belonging to the service provider. Two of the resources are transmission channels for the buyers to choose (i.e., transmission capacity) indexed as $j = 1$ and $j = 2$. The other two resources are servers (i.e., commoditized computational capacity), indexed as $j = 3$ and $j = 4$. Therefore, there are two complementary groups in the service provider, each of the groups contains two substitutable resources, i.e., $G = 2$, and $M_1 = M_2 = 2$ as shown in Fig. 1(b).

5.1 Example 1: A Basic Case

First we consider a basic case where buyers' preferences on either complementary or substitutable resources are zero. That is, $\forall i, l_i(t_i) \equiv h_i(t_i) \equiv 0$. The optimal mechanism (p^*, c^*) as in (12) and (11) reduces to

$$p^* = \arg\max_p \sum_{i=1}^{2} \mathbb{E}_{\mathbf{t}} \Big\{ \sum_{j=1}^{4} \big(v_{ij}(t_i) - \frac{1-F(t_i)}{f(t_i)} \frac{\partial v_{ij}(t_i)}{\partial t_i}\big) p_{ij}(t_i) \Big\}$$

$$c_i^*(t_i) = \mathbb{E}_{\mathbf{t}} \Big\{ \sum_{j=1}^{4} p_{ij}^*(t_i) v_{ij}(t_i) - \sum_{j=1}^{4} \int_{\underline{t_i}}^{t_i} \frac{\partial v_{ij}(x)}{\partial t_i} p_{ij}^*(x) dx \Big\}.$$

subject to (8), (9) and (10). As a result, the auction becomes a general single-commodity auction case as in [6].

5.2 Example 2: A Symmetric Case

We consider a two-buyer, symmetric scenario as follows:

- In this symmetric case, we discuss the situation that the buyers' valuations on the first group of resources ($j = 1, 2$) are linear, i.e., $v_{i1} = \gamma_{i1} t_i$ and $v_{i2} = \gamma_{i2} t_i$. The expressions of v_{i1} and v_{i2} mean that user i's valuation on a communication channel (resources 1 and 2) is linear with respect to the type. The relation between valuation functions and types indicates that the valuation linearly increases as the buyers' preference to the resources (transmission bandwidth) increases.
- The buyers' valuations on the second group of resources ($j = 3, 4$) are increasing and concave functions of types ($\frac{\partial f(x)}{\partial x} \geq 0$ and $\frac{\partial^2 f(x)}{\partial x^2} \leq 0$), where the marginal payoffs decrease as the types increase. We consider $v_{i3} = \log(1 + \theta_{i3}/t_i)$ and $v_{i4} = \log(1 + \theta_{i4}/t_i)$, $i = 3, 4$. Similarly, the type t_i in v_{i3} and v_{i4} represents user i's preference on computational capacity (i.e., servers). The intuition behind such concave valuations is that the corresponding resources might not be the main bottleneck of performance. As a result, the buyers still prefer higher level resources but with less desire.
- User's type t_i is uniformly distributed from 0 to 1, for $F(t_i) = t_i$ and $f(t_i) = 1$. Also, the lower bound of t_i is $\underline{t_i} = 0$ and upper bound of t_i is $\overline{t_i} = 1$.

The optimal resource allocation p^* as in (11) is optimized as follows:

$$p^* = \arg\max_p \mathbb{E}_{\mathbf{t}} \Big\{ \sum_{i=1}^{2} \sum_{j=1}^{4} \big[v_{ij}(t_i) + (t_i - 1)\frac{\partial v_{ij}(t_i)}{\partial t_i}\big] p_{ij}(t_i) + \sum_{i=1}^{2} h_i \big[p_{i1}(t_i) + p_{i2}(t_i)\big]\big[p_{i3}(t_i) + p_{i4}(t_i)\big] - \sum_{i=1}^{2}\sum_{j=1}^{4} l_i(t_i) p_{ij}(t_i) \Big\}. \tag{14}$$

From the allocation scheme defined in (14), we can see that the service provider tends to adopt combinatorial allocation as the dominant strategy. That is, to maximize the revenue, the service provider will allocate the whole group of resources to a buyer, and another group of resources as a whole to another buyer.

5.3 Example 3: An Asymmetric Case

A simple asymmetric auction scenario is considered. The basic parameter setting is the same as that of the symmetric scenario, except the users' valuation functions are heterogeneous, as follows:

- The valuation functions of user 1 are strictly concave. We assume in this case that $v_{11} = \gamma_{11}\sqrt{t_1}$, $v_{12} = \gamma_{12}\sqrt{t_1}$, $v_{13} = \sqrt{t_1}\log(1 + \theta_{13}/\sqrt{t_1})$ and $v_{14} = \sqrt{t_1}\log(1 + \theta_{14}/\sqrt{t_1})$.
- The valuation functions of user 2 are different from that of user 1, i.e., $v_{21} = \gamma_{21}t_2$, $v_{22} = \gamma_{22}t_2$, $v_{23} = t_2\log(1+\theta_{23}/t_2)$ and $v_{24} = t_2\log(1+\theta 24_b/t_2)$. We set in our examples that $\forall i \in \{1, 2\}$, $\gamma_{i1} = 10$, $\gamma_{i2} = 20$, $\theta_{i3} = 1.25 \times 10^6$ and $\theta_{i4} = 3.75 \times 10^6$.
- The distribution of type t_1 is not uniform, i.e., $F(t_1) = (t_1)^2$ and $f(t_1) = 2t_1$, where $t_1 \in [0, 1]$, while type t_2's distribution is still uniform over $[0, 1]$.

We fix user 2's type t_2 at 0.6, and gradually increase user 1's type t_1 from 0 to 1. As shown in Figs. 2(a) and (b), when t_1 is small, user 2 is the winner and contributes a flat revenue to the service provider. When t_1 increases, user 1's valuation may surpass that of user 2. User 1 becomes the winner after a turning point (0.5 in this case). The decreasing of marginal utility causes the user 1's utility to decrease as in Fig. 2(a). Also, derived from (13), the utility of user 1 is convex after the turning point as shown in Fig. 2(a).

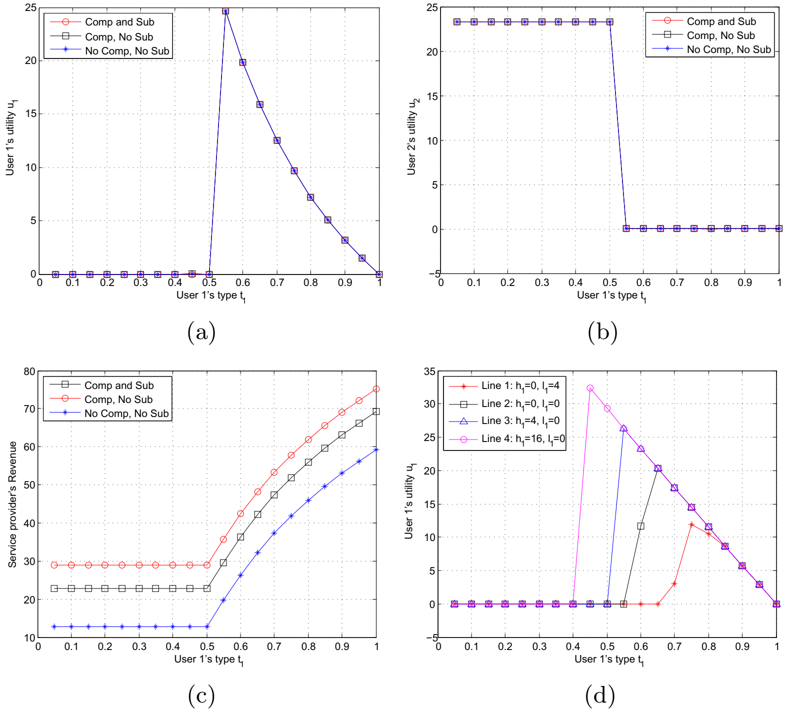
5.4 Impacts of Premium and Discount Factors

Fig. 2(c) resulted from the asymmetric scenario (Section 5.3) depicts the effects of complementary and substitutable resources to the revenue of the cloud service provider (i.e., seller). It is shown that the buyers will have extra utility on the complements, so that the service provider's revenue will increase according to (11). Also, the existence of substitutable resources will decrease the total revenue.

In the examples, the premium and discount factors of users can be set in 3 different cases as follows:

- With both complementary and substitutable resources: $h_1 = h_2 = 4$ and $l_1 = l_2 = 1.5$. In this case, the resources in the same group are substitutable for the users with a discount factor of 1.5, and resources from different groups are complementary with a premium factor of 4.
- With complementary resources only: $h_1 = h_2 = 4$ and $l_1 = l_2 = 0$. The resources from different groups are complementary. However, the resources from the same group are not substitutable.
- Without any complementary and substitutable resources: $h_1 = h_2 = 0$ and $l_1 = l_2 = 0$.

From Fig. 2(d) we can compare the auction results given the different settings of premium and discount factors. It is clear that the system with only complementary resources has the optimal performance in terms of utilities of users and revenue of the service provider, compared with other cases when other settings



(a) (b)

(c) (d)

Fig. 2. Asymmetric scenario: (a) user 1's utility to her type t_1, (b) user 2's utility to user 1's type t_1, and (c) the service provider's revenue to user 1's type t_1. Premium and discount factors' impacts: (d) user's utility under different settings of premium and discount factors.

are the same. This result is intuitive, since the premium factor of the complementary resources increases the revenue of the service provider according to (11). However, for the system with both complementary and substitutable resources, and the system without any complementary and substitutable resources, the results may vary, since the discount factor might degrade the performance.

For a user, the premium and discount factors can affect the resource allocation done by the service provider. According to (11), the premium factor increases the total revenue, so the service provider is in favor of allocating resources to the user with a higher premium factor. For the discount factor, the effect is just opposite. Fig. 2(d) shows the impact of premium and discount factors to the user 1's utility. In this case, user 2's type t_2 is fixed at 0.6, and has the premium and discount factors set as $h_2 = l_2 = 0$. Then, t_1 increases from 0 to 1. The parameter setting is the same as that used in the symmetric scenario except h_i and l_i. Line 2 (see legend of Fig. 2(d)) is a reference result for $h_1 = l_1 = 0$. It is shown by the figure that, a higher premium factor will increase the user's utility, and thus the total revenue is increased. The service provider may allocate the resources to the user even the type of the user is relatively low. Similarly, the existence of high discount factor will decrease the user's chance of being allocated with the resources.

6 Summary

In this paper, we have addressed an analytical auction model for the resource allocation problem in a mobile cloud computing system. The complementary and substitutable cloud resources have been considered. We have analyzed the model, and solved the optimization problem to maximize the revenue of the service provider given the proposed auction mechanism. From the numerical results, we have found the changing of users' utilities and the service provider's revenue with respect to users' types, the optimal allocation schemes of the service provider, and the impacts of users' premium and discount factors to the users' utilities and the service provider's revenue.

References

1. http://aws.amazon.com/ec2
2. Li, Z., Wang, C., Xu, R.: Computation offloading to save energy on handheld devices: A partition scheme. In: Proceedings of the 2001 International Conference on Compilers, Architecture, and Synthesis for Embedded Systems (CASES), pp. 238–246 (2001)
3. Armbrust, M., Fox, A., Griffith, R., et al.: A view of cloud computing. Communications of the ACM 53, 50–58 (2010)
4. Wang, Q.: Mobile cloud computing. Master of science thesis, University of Saskatchewan, Canada (2011)
5. Hoang, D.T., Lee, C., Niyato, D., Wang, P.: A survey of mobile cloud computing: Architecture, applications, and approaches. In: Wireless Communications and Mobile Computing, doi:10.1002/wcm.1203 (in press)
6. Myerson, R.B.: Optimal auction design. Mathematics of Operations Research 6(1), 58–73 (1981)
7. Branco, F.: Multiple unit auctions of an indivisible good. Economic Theory 8(1), 77–101 (1996)
8. Levin, J.: An optimal auction for complements. Games and Economic Behavior 18(2), 176–192 (1997)
9. Burguet, R.: The condominium problem; auctions for substitutes. Review of Economic Design 9(2), 73–90 (2005)
10. Parsons, S., Rodriguez-Aguilar, J.A., Klein, M.: Auctions and bidding: A guide for computer scientists. ACM Computing Surveys 43, 10:1–10:59 (2011)
11. Mullins, C.L.: Real-time auction of cloud computing resources. U.S. Patent 2010/0076856 A1 (March 25, 2010)
12. Lin, W.-Y., Lin, G.-Y., Wei, H.-Y.: Dynamic Auction Mechanism for Cloud Resource Allocation. In: 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGrid), pp. 591–592 (May 2010)
13. Zhang, Y., Niyato, D., Wang, P.: An Auction Mechanism for Resource Allocation in Mobile Cloud Computing Systems, http://arxiv.org/abs/1304.6176

A Novel Delay-Resilient Remote Memory Attestation for Smart Grid

Xiaofei He[1], Xinyu Yang[1], Rui Li[1], and Qingyu Yang[2]

[1] Department of Computer Science and Technology, Xi'an Jiaotong University,
Xi'an, 710049, P.R. China
{hexiaofei,liruixjtu}@stu.xjtu.edu.cn, yxyphd@mail.xjtu.edu.cn

[2] The School of Electronic and Information Engineering, Xi'an Jiaotong University,
Xi'an, 710049, P.R. China
yangqingyu@mail.xjtu.edu.cn

Abstract. Smart measurement devices play an important role in smart grid and might always be connected through open network interfaces. In this scenario, the adversary could launch code injection attacks to compromise these measurement devices and gain benefits by these compromised devices. To deal with this issue, a number of attestation schemes have been designed to defense the malicious attacks in the past. However, because the detection methods of these schemes are based on extra CPU clock cycles, they could be ineffective when the network delivery delay is significant. To address this problem, in this paper we propose a novel Delay-resilient Remote Memory Attestation scheme (DRMA), which can eliminate the impact of network delivery delay in the multi-hop networks and achieve great accuracy on compromised measurement devices detection. Specially, without sending beacon packets periodically, the proposed scheme can not only get the real-time end-to-end delay via evaluating the time difference reported by the relay nodes in the challenge-response attestation process, but also reduce the network load and achieve great accuracy of network delay. Via extensive theoretical analysis and experiments, our scheme shows better performance and less computing overhead in comparison with existing schemes.

Keywords: Smart measurement devices, Code injection attack, Delay-resilient memory attestation, Smart grid.

1 Introduction

With the development of information technology, numerous countries devote to the smart grid research and evolving the traditional grid into the smart grid gradually [1]. Smart grid, as the next generation of the power grid can monitor and control the state of power grid with more secure, reliable, economic and efficient [2]. In particular, smart grid deployed with smart measurement devices such as smart meters, can not only enhance the interactivity between utilities and customers, but also can achieve great efficiency on power transmission and distribution. However, Because those measurement devices may connected through

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 88–99, 2013.

open network interfaces, the adversary could launch code injection attacks to compromise and capture these measurement devices [3]. Using these compromised devices, the adversary could launch malicious attacks to inject false energy demand information to disrupt the generation dispatch [4] and electricity market operations [5] in smart grid.

To address this issue, a lot of efforts have been developed to detect compromised smart measurement devices in smart grid [6,7,8]. However, the existing attestation schemes verify remote devices based on extra CPU clock cycles caused by memory forgery attacks [9]. If the network delay is significant, these schemes could be ineffective. Unfortunately, most of the network devices are deployed far from the verifier in the smart grid [10], and the large end-to-end delivery delay in retransmission process makes it hard to distinguish the increased computing time delay caused by memory forgery attacks from normal network delay. Furthermore, the end-to-end delay will also increase when the network has poor quality or plenty of traffic, which leads more difficult to determine whether the remote device is compromised. In addition, the adversary might launch Denial of Service (DoS) attacks [11] or selective forwarding attacks [12] to disrupt the network performance, which leads the verifier cannot verify the correctness of memory of the remote devices. Hence, the existing schemes could not exactly detect out the compromised measurement devices in the grid with great end-to-end delivery delay, because their detection is based on the actual response time without considering the real-time delivery delay of the grid.

In this paper, we propose a novel delay-resilient remote memory attestation scheme (DRMA), which can detect the compromised measurement devices based on the response time with considering the real-time delivery delay of the grid. In particular, we evaluate the delay via using the time difference reported by the relay nodes in the challenge-response attestation process. Based on the obtained real-time delivery delay, our proposed scheme can eliminate the impact of delivery delay in the multi-hop networks and exactly detect the compromised devices in the grid with great end-to-end delivery delay. In addition, DRMA does not need to send beacon packets periodically, which would lead the reduction of network load in the grid. In brief, our proposed DRMA can not only achieve great accuracy on detection of compromised smart measurement devices in the grid with great end-to-end delivery delay, but also reduce the network load and achieve low computing overhead.

Via extensive theoretical analysis and simulation experiments, we evaluate the security of DRMA in comparison with the existing approaches in terms of response time, accuracy on compromised devices detection and computing overhead. Our data show that DRMA achieves better performance than existing schemes. For example, the response time of DRMA is always shorter than that of existing schemes. When the forwarded hops reach to 20, DRMA still has great accuracy on compromised devices detection while the other schemes cannot distinguish the increased computing time from the network delay. In the same situation, the computing overhead of DRMA is one order of magnitude smaller than that of the others.

The remainder of this paper is organized as follows: We introduce the Advanced Metering Infrastructure (AMI) structures and the assumptions in Section 2. We present our proposed scheme in Section 3. In Section 4, we analyze the security properties of our scheme. In Section 5, we show experimental results to validate our theory and findings. We review related work in Section 6 and conclude this paper in Section 7.

2 Preparation

In this section, we first introduce the structure of Advanced Metering Infrastructure (AMI), and then describe all assumptions in our paper.

2.1 Advanced Metering Infrastructure

Smart grid, as shown in Fig. 1, includes two types of lines, one is the power line which is used for power transmission and distribution, and the other one is the communication line which is used to connect the smart measurement devices in smart grid and transmit the interaction information between customers and utilities.

As the communication network is merged into smart grid, a novel advanced metering infrastructure (AMI) also be developed in smart grid and used to measure, collect, store and analyze the customers' electricity usage information, as shown in Fig 2. The smart meter is a kind of sensors that records the real-time (or quasi real-time) measurement values, such as voltage, current, power, consumption and the other data. Data collectors (DC) send the data sensed by smart meters to the control center through the ad hoc network of smart meters. The control center processes and stores the data collected by DCs. With AMI, customers could be exchange the electricity usage information with control center and also can participate the grid operation via controlling their power supply and consumption by smart meters. In this way, utilization of power transmission and distribution will be maximum.

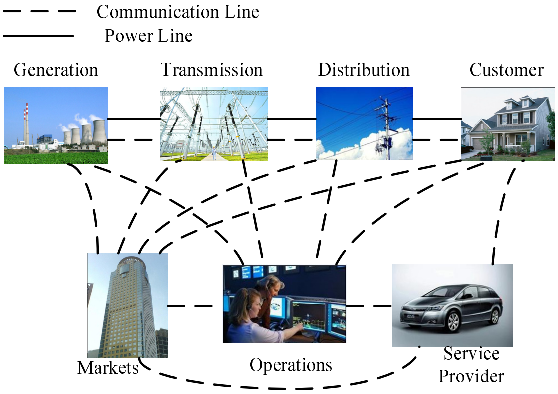


Fig. 1. The Structure of the Smart Grid

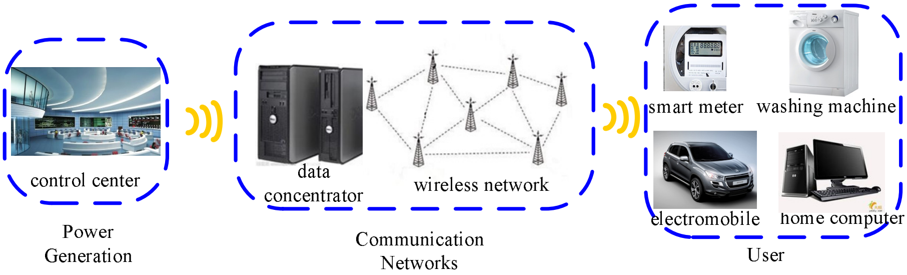


Fig. 2. The Structure of the Smart Grid AMI

2.2 Assumption

In our scheme, we assume that the verifier in smart grid is well protected and could not be compromised by adversary, and all information stored in verifier could not be leaked to adversary. Because the compromised device will be detect out by the returned time of checksum in our scheme, the verifier should understand the hardware structure and software information of the remote devices, including the CPU clock speed, the memory architecture, and the instruction set architecture (ISA) of the microcontroller, etc, when the network equipment is deployed. To make the verifier obtain the correct time of computing the memory checksum of devices, the verifier should have a local copy which stores the memory contents of the remote devices. In addition, we also assume that the verifier can activate the verification procedure on the device remotely, which has access to its memory contents.

Because all smart measurement devices may be connected through open network interfaces, the adversary could launch code injection attacks to compromise or capture these measurement devices and take full control of the compromised devices. Using the compromised devices, the adversary can change their memory contents. However, because the compromised devices are usually integrated devices, it is extremely difficult for adversary to change their hardware. Moreover, in smart grid it is extremely difficult to replace the network equipment without interrupting its normal operation. Therefore, all assumptions in our scheme are easily to be realized in realities.

3 Our Approach

3.1 Basic Idea

To make the attestation scheme effectively detect out the compromised smart measurement devices in the grid with great end-to-end delivery delay, in this paper we propose a novel delay-resilient remote memory attestation scheme (DRMA), which can evaluate the real-time network delay via using the time difference reported by relay nodes in the process of the challenge-response protocol. The basic idea of our scheme is described below. As shown in Fig. 3, DRMA consists of the following steps.

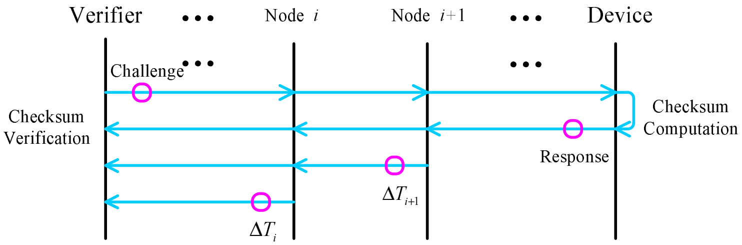


Fig. 3. The attestation process of DRMA

3.2 Challenge Generation

The challenge request will be sent from the verifier to the remote device to verify whether the memory is changed by the adversary. To avoid an adversary launching the replay attacks to reduce the computing time for current received challenge message, i.e., returning the old response message to response the current challenge message from the verifier, a random challenge message is needed. Hence, the challenge message generation has three steps:

1) The verifier generates a random key based on current time of the verifier.

$$Challenge : key = hash(T_{Verifier}) \tag{1}$$

2) The message is generated to be sent.

$$m = \{key \| W_k\} \tag{2}$$

3) The verifier sends a challenge message to the remote device with the corresponding Message Authentication Code $MAC_n(m)$ and records when the packet is sent.

$$Verifier \rightarrow n : request = \{m \| MAC_n(m)\} \tag{3}$$

$$Ts_{Verifier} = Tcurrent_{Verifier} \tag{4}$$

3.3 Challenge Transmission

The challenge transmission of the relay nodes involves two steps:

1) The relay node forwards the challenge to the next hop.

$$i \rightarrow i+1 : request \tag{5}$$

2) The relay node records the time when the request is forwarded.

$$Ts_i = Tcurrent_i \tag{6}$$

3.4 Checksum Generation

The remote device calculates the checksum in the following steps:

1) The remote device gets the key from the verifier and initiates the RC4 to generate the random memory addresses.

$$A_k = RC4(key, W_k) \tag{7}$$

2) The remote device gets the memory contents of the addresses to calculate the checksum.

$$C_k \leftarrow Mem[A_k] \oplus C_{k-1} + RC4_k \tag{8}$$

$$C_k \leftarrow \ rotate\ left\ one\ bit(C_{k-1}) \tag{9}$$

3) The remote device generates the checksum and returns it to the verifier with the corresponding MAC.

$$C = \{C_1 \| C_2 \| \cdots \| C_k\} \tag{10}$$

$$response = \{C \| MAC_n(C)\} \tag{11}$$

3.5 Checksum Transmission

The checksum transmission of the relay nodes involves the following steps:

1) The relay node records the time when the response received.

$$Tr_i = Tcurrent_i \tag{12}$$

2) The relay forwards the response to the next hop.

$$i \rightarrow i - 1 : response \tag{13}$$

3) The relay node calculates the time difference between the request and the response.

$$\varDelta T_i = Tr_i - Ts_i \tag{14}$$

4) The relay node reports the time difference to the verifier after a while.

$$m = \{\varDelta T_i \| MAC_i(\varDelta T_i)\} \tag{15}$$

3.6 Checksum Verification

The process to verify the response of the remote device is as follows:

1) The verifier records when the response is received.

$$Tr_{Verifier} = Tcurrent_{Verifier} \tag{16}$$

2) The verifier computes the $MAC'_n(C)$ separately, and compares it with the $MAC_n(C)$ in the response.

$$Compare(MAC_n(C), MAC'_n(C)) \tag{17}$$

3) If the MAC are the same, the verifier will compute checksum C' separately and record the computing time $T'_{checksum}$.

$$C' = \{C'_1 \| C'_2 \| \cdots \| C'_k\} \tag{18}$$

4) The verifier compares the checksum C' with the C in the response.

$$Compare(C, C') \tag{19}$$

3.7 Determination

The process of the determination consists of the following steps:

1) If the checksums are the same, the verifier will calculate the time spent in the entire challenge response.

$$\varDelta T_{Verifier} = Tr_{Verifier} - Ts_{Verifier} \tag{20}$$

2) The verifier receives the reports from the relay nodes, and computes the delay on each hop along the path.

$$\begin{cases} D_1 = \frac{\varDelta T_1 - \varDelta T_{Verifier}}{2} \\ D_i = \frac{\varDelta T_i - \varDelta T_{i-1}}{2} \qquad (1 < i \le n-1) \end{cases} \tag{21}$$

3) After eliminating the exceptional data by Pauta criterion, the average delay in the network is estimated by the sample mean.

$$\overline{delay} = \frac{\sum\limits_i D_i}{n'} \qquad (\mid D_i - \bar{D} \mid \le 3s) \tag{22}$$

where n' is the number of samples after eliminating the exceptional data, $\bar{D}$ is the sample mean.

4) According to the calculated transmission delay, the verifier gets the time spent on computing checksum.

$$T'_{checksum} = \varDelta T_{Verifier} - n \cdot \overline{delay} \tag{23}$$

5) Comparing it with the normal time consumed by the process of the checksum generation $T_{checksum}$, the verifier can determine whether the memory of the remote device is changed by an adversary.

$$Compare(T_{checksum}, T'_{checksum}) \tag{24}$$

4 Security Analysis

4.1 Local Attack

An adversary might modify the memory on the remote node. the changed memory contents on the node, however, will lead the failure of the checksum computation. In this way, the compromised devices cannot generate the correct checksum, and thus they could be accurately detected out by the detection of "challenge - response" mechanism. In our proposed scheme, because the checksum has n bits, the probability of coincidence is only $\frac{1}{2^n}$.

An adversary might also modify the attestation procedure to compute the checksum accurately. The extra `if` statement will divert a position in the memory where store the original values, but it will cause a detectable delay. The delay is related to the number of memory access of the verification procedure. Generally, the execution times can be set to:

$$N = \alpha \frac{delay}{t} \tag{25}$$

where *delay* is the end-to-end delay in the networks, t is the increased computation time caused by an extra `if` statement, α is the scale factor. The greater α is, the longer the computing time spent on checksum is. Generally, α can be set to 10˜100 in order to ensure that the increased delay is one or two order of magnitude larger than the network delay, which can be detected by the challenge-response mechanism.

4.2 Network Attack

Besides node memory code injection attacks, the other attacks against the challenge-response attestation process will also exist in the networks.

Relay Nodes Report False Time Difference. An adversary may attempt to exploit the compromised node to disturb the delay estimation process. However, the time difference reported by relay nodes should be a sequence as follows:

$$\Delta T_1, \Delta T_2, \ldots, \Delta T_{n-1} \tag{26}$$

The delay between any two adjacent nodes is:

$$\begin{cases} D_1 = \frac{\Delta T_1 - \Delta T_{Verifier}}{2} \\ D_i = \frac{\Delta T_i - \Delta T_{i-1}}{2} \qquad (1 < i \leq n-1) \end{cases} \tag{27}$$

According to the central limit theorem, this delay approximates a normal distribution:

$$D_i \sim N(\mu, \sigma^2) \tag{28}$$

where expectation μ is estimated by the sample mean $\bar{D}$, variance σ^2 is estimated by the sample variance s^2.

If the relay node i sends false time difference data, the time difference between adjacent nodes will significantly offset the sample mean.

$$\mid D_i - \bar{D} \mid > 3\sigma \text{ or } \mid D_{i+1} - \bar{D} \mid > 3\sigma \tag{29}$$

Thus the verifier could remove the false data out.

Relay Nodes Collusion. An adversary may also attempt to disrupt the normal delay estimation process, using the adjacent compromised node to launch a collusion attack. To cover up the computing time at the destination node, collusion compromised nodes must be adjacent to it. For an n-hop path, if the attacker takes control of m consecutive nodes adjacent to the destination node, the increased delay could be divided into m parts to add to the reports of the compromised nodes. In order to avoid being eliminated by the Pauta criterion, the time difference reported by the compromised nodes should be:

$$\Delta' T_{n-m+i} = \Delta T_{n-m+i} - i \cdot \frac{2\alpha T_{if}}{m} \qquad (1 \le i \le m) \tag{30}$$

Where the T_{if} is the increased computing time caused by the extra `if` statements. α is the scale factor.

Further, the changed the delay between the adjacent nodes is

$$D'_i = \frac{\Delta T'_i - \Delta T'_{i-1}}{2} = D_i + \frac{\alpha T_{if}}{m} \qquad (n - m + 1 \le i \le n) \tag{31}$$

To avoid being as outliers, the changed network delay should meet:

$$\begin{aligned} &\min && m \\ &s.t. && \begin{cases} \mid D'_i - \overline{D}' \mid \le 3\sigma & (n - m + 1 \le i \le n) \\ \mid D_i - \overline{D} \mid \le 3\sigma & (1 \le i \le n - m) \end{cases} \end{aligned} \tag{32}$$

where the $\bar{D}$ is normal sample mean $\bar{D} = \frac{\sum_i D_i}{n}$, $\bar{D}'$ is the sample mean after some reports is changed $\bar{D}' = \frac{\sum_{i=0}^{n-m} D_i + \sum_{i=n-m+1}^{n} D'_i}{n} = \bar{D} + \frac{\alpha T_{if}}{n}$.

Therefore we can get:

$$m \ge \frac{1}{\frac{3\sigma}{\alpha T_{if}} + \frac{1}{n}} \tag{33}$$

Thus, with the increase of the number of hop n in the path, the number of compromised nodes m which the adversary should control also increases. If $n = 10$, m is about 5 (σ is 0.001,T_{if} is 0.3, α is 0.5).

The farther the remote node is from the verifier, the more nodes the attacker needs to control, and the more difficult to launch an attack.

5 Evaluation

5.1 Evaluation Setup

We use ns-3 in our simulation. In the experiments, we simulate a static wireless ad hoc network consists of 225 nodes. All nodes uniformly distributed in a grid of 15*15. All the nodes are the same, with the 802.11b NICs in ad hoc mode. The verifier is located at (0, 0), which verifies the nodes on the different positions in the net-work respectively. The network has an 802.11b physical layer with DsssRate1Mbps and AODV route protocol.

5.2 Evaluation Result

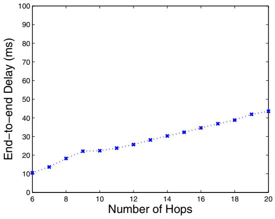


Fig. 4. The end-to-end delay vs. number of hops

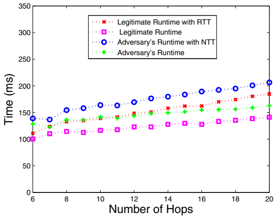


Fig. 5. The response time vs. number of hops. (Low overhead)

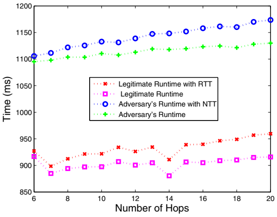


Fig. 6. The response time vs. number of hops. (High overhead)

The End-to-end Delay Vs. Number of Hops. As shown in Fig. 4, with the increase of the number of hops, the end-to-end delay in the network is gradually increased, becoming more significant.

The Response Time Vs. Number of Hops. As shown in Fig. 5, with the increase of the number of hops, the verification time that a verifier spends for once is also gradually increasing. The verification procedure which an adversary modifies will cost more computing time than the normal. Normal computing time is approximately 100 to 150ms, which is difficult to distinguish with the network delay.

As shown in Fig. 6, the increased number of the memory access will make the time curve separate from the normal curve. However, it will significantly make the computing time increase to several times of previously time.

6 Related Work

In this section, we review related work in the area of the remote device attestation.

SCUBA[6] proposed a protocol for the detection and recovery node sensor network is captured, according to the challenge-response time to determine whether it should send an update patch. The OMAP [7] proposed a one-way transmission protocol, where the remote device sends the time and the corresponding checksum to the verifier to verify the correctness. SWATT [8] show that the extra instructions which an attacker injects will cause a detectable slow down. CAK [13] proposed a cumulative attestation to detect attacks which one might call Time-Of-Use-To-Time-Of-Check (TOUTOC) inconsistencies. SAKE [14] presents a protocol for establishing a shared key with secrecy and authenticity between any two neighboring nodes. Pioneer [15] also is a similar code attestation on untrusted legacy hosts. Yi Yang et al. [16] proposed two distributed software-based attestation schemes based on a pseudorandom noise generation mechanism and a lightweight block-based pseudorandom traversal algorithm. A. Seshadri et al. proposed FIRE [17] to detect and repair compromised nodes.

all the above did not take into account any impact of the network delay. Thus an adversary might launch attacks on the end-to-end delay in the networks to interrupt the detection of the normal attestations. Our work will mitigate the threats.

7 Conclusion

In this paper, we present a novel delay-resilient remote memory attestation scheme (DRMA) by taking into account the impact of network transmission delay so that this scheme has good flexibility to the increased delay caused by the growth of the number of hops in the wireless multi-hop networks or cyber attacks. In DRMA, each relay node will send the time difference between the challenge request and the response to the verifier, and then the verifier calculates the network delay using the time difference reported by the relay nodes. Via both theoretical analysis and simulation experiments , our data show that our scheme can significantly reduce the computational overhead of the challenge-response attestation process and obtain a faster response, which makes it have a higher degree of adaptability in the smart grid AMI.

References

1. Li, F., Qiao, W., Sun, H., Wan, H., Wang, J., Xia, Y., Xu, Z., Zhang, P.: Smart transmission grid: Vision and framework. IEEE Transactions on Smart Grid 1(2), 168–177 (2010)
2. DeBlasio, R., Tom, C.: Standards for the smart grid. In: Energy 2030 Conference, 2008, pp. 1–7. IEEE (2008)

3. Huang, Y., Esmalifalak, M., Nguyen, H., Zheng, R., Han, Z., Li, H., Song, L.: Bad data injection in smart grid: attack and defense mechanisms. IEEE Communications Magazine 51(1), 27–33 (2013)
4. Yang, X., Lin, J., Moulema, P., Yu, W., Fu, X., Zhao, W.: A novel en-route filtering scheme against false data injection attacks in cyber-physical networked systems. In: 2012 IEEE 32nd International Conference on Distributed Computing Systems (ICDCS), pp. 92–101. IEEE (2012)
5. Xie, L., Mo, Y., Sinopoli, B.: Integrity data attacks in power market operations. IEEE Transactions on Smart Grid 2(4), 659–666 (2011)
6. Seshadri, A., Luk, M., Perrig, A., van Doorn, L., Khosla, P.: Scuba: Secure code update by attestation in sensor networks. In: Proceedings of the 5th ACM Workshop on Wireless Security, pp. 85–94. ACM (2006)
7. Song, K., Seo, D., Park, H., Lee, H., Perrig, A.: Omap: One-way memory attestation protocol for smart meters. In: 2011 Ninth IEEE International Symposium on Parallel and Distributed Processing with Applications Workshops (ISPAW), pp. 111–118. IEEE (2011)
8. Seshadri, A., Perrig, A., Van Doorn, L., Khosla, P.: Swatt: Software-based attestation for embedded devices. In: Proceedings. 2004 IEEE Symposium on Security and Privacy, pp. 272–282. IEEE (2004)
9. Castelluccia, C., Francillon, A., Perito, D., Soriente, C.: On the difficulty of software-based attestation of embedded devices. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, pp. 400–409. ACM (2009)
10. Hart, D.G.: Using ami to realize the smart grid. In: Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, 2008, pp. 1–2. IEEE (2008)
11. Lu, Z., Lu, X., Wang, W., Wang, C.: Review and evaluation of security threats on the communication networks in the smart grid. In: Military Communications Conference, MILCOM 2010, pp. 1830–1835 (2010)
12. Bysani, L., Turuk, A.: A survey on selective forwarding attack in wireless sensor networks. In: 2011 International Conference on Devices and Communications (ICDeCom), pp. 1–5 (2011)
13. LeMay, M., Gunter, C.A.: Cumulative attestation kernels for embedded systems. IEEE Transactions on Smart Grid 3(2), 744–760 (2012)
14. Seshadri, A., Luk, M., Perrig, A.: Sake: Software attestation for key establishment in sensor networks. In: Nikoletseas, S.E., Chlebus, B.S., Johnson, D.B., Krishnamachari, B. (eds.) DCOSS 2008. LNCS, vol. 5067, pp. 372–385. Springer, Heidelberg (2008)
15. Seshadri, A., Luk, M., Shi, E., Perrig, A., van Doorn, L., Khosla, P.: Pioneer: verifying code integrity and enforcing untampered code execution on legacy systems. In: ACM SIGOPS Operating Systems Review, vol. 39, pp. 1–16. ACM (2005)
16. Yang, Y., Wang, X., Zhu, S., Cao, G.: Distributed software-based attestation for node compromise detection in sensor networks. In: 26th IEEE International Symposium on Reliable Distributed Systems, SRDS 2007, pp. 219–230. IEEE (2007)
17. Seshadri, A., Luk, M., Perrig, A., Doorn, L.V., Khosla, P.: Using fire & ice for detecting and recovering compromised nodes in sensor networks. Tech. rep., DTIC Document (2004)

A Source-Relay Selection Scheme with Power Allocation for Asymmetric Two-Way Relaying Networks in Underground Mines

Song Li[1], Yanjing Sun[1], and Rufei Ma[2]

[1] School of Information and Electrical Engineering,
China University of Mining and Technology, Xuzhou, China
[2] School of Electronics and Information Engineering, Tongji University, Shanghai, China

Abstract. In this paper, the two-way opportunistic amplify-and-forward relaying system with multi-source multi-relay in underground tunnel is investigated; in which two sources intend to exchange information with help of one relay. The joint source-relay selection with power allocation is proposed to minimize the total transmit power subject to constraints on the received signal-to-noise-ratio (SNR) of each source node. We show that this problem has a closed-form solution and requires only a few parameters to be broadcasted to all relays during the source-relay selection period. Simulation results show that the outage probability can be substantially decreased using the proposed power allocation scheme when two source node have different quality of service (QoS) requirement, comparing to existing power allocation algorithms.

Keywords: two-way relay networks, source-relay selection, power allocation, outage probability.

1 Introduction

Reliable and efficient communication networks are needed to improve the safety and productivity in underground mines. Diverse information stream are delivered in wireless underground tunnels, including video steam, audio steam and control information stream [1]. Two-way relay communications have recently attracted considerable interest, and transmission schemes in two-way relay networks have been analyzed wildly [2][3]. Two-way relaying is a spectral efficient protocol when two sources need to exchange information with the help of one or multiple relays [4]. Based on complexity criterion, the amplify-and-forward (AF) protocol is the simplest strategy because it only requires an amplification processing at the relays. The transmission in amplify-and forward two-way relay network takes place in two time slot. Two transceivers first transmit at the same time to one or multiple relays. The relay receives a superimposed signal, then amplifies the received signal, and forwards it back to both transceivers [5].

In two-way opportunistic relay (TWOR) networks, the performance of wireless relay networks can further be enhanced by properly selecting the relays for transmission [6]-[8]. In [6], relay selection methods were reported for conventional one-way AF

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 100–109, 2013.

schemes to achieve full spatial diversity. In [7], the authors presented a max–min signal-to-noise ratio (SNR)-based relay selection algorithm for two-way relay networks. In [8], the author presented a relay selection to minimize the system symbol error rate (SER). Consequently, it is beneficial to design an effective relay selection scheme for the coherent bidirectional transmission scheme with multiple relays and to achieve spatial diversity.

Power allocation in one-way relay systems has intensively been studied. Because two-way relay system works quite differently and is more complex than one-way relay system, the power allocation algorithms developed for one-way relaying cannot readily be used in two-way relay systems. Most work on power allocation for two-way relaying, e.g., [9], was proposed to maximize the sum rate of the user pair. In [10], the authors consider power allocation with wireless network coding in a multiple-relay, multiple-user networks using convex optimization. In [11], power allocation strategies are proposed to maximize the sum rate and the diversity order, respectively. In [12], two power allocation algorithms were proposed to maximize the upper bound of the average sum rate and to achieve the tradeoff of outage probability between two transceivers, respectively. The power allocation algorithm proposed in [13] aims to maximize the smaller of the two transceiver SNR. But the algorithm in [13] did not consider the scenario that two transceivers have different QoS requirement, e.g., different transmit data rate requirement and different symbol error ratio requirement.

We herein present a power allocation scheme in two-way opportunistic relay networks with multiple source nodes and multiple relay nodes to minimize the total transmit power subject to constraints on two transceivers' received SNRs. We obtain the closed-form solution of optimal power allocation for each node. Then the source-relay pair with the minimum total transmit power is selected. The source-relay selection can be operated either by centralized manner or distributed manner. The performance of the proposed scheme is verified through simulations.

The rest of this paper is organized as follows. Section 2 introduces system model of the two-way relay system and presents the optimization objective function of the joint source-relay selection and power allocation problem. In Section 3, We present our proposed power allocation scheme and source-relay selection seperately. In Section 4, the outage probability is analyzed when the proposed power allocation scheme and source-relay selection scheme is adopted. Simulation results are illustrated in Section 5 and conclusions are presented in Section 6.

2 System Model and Problem Formulation

2.1 System Model

We consider a two-way relaying system, in which one fixed source terminal A intends to exchange information with one out of M mobile source terminals B_m (m=1, 2,…,M) with the help of one out of N relay nodes R_n (n=1, 2, …, N).

We assume no direct communication between the two source nodes because of the poor quality of the channel between them. All nodes are equipped with a single antenna and operate in half-duplex mode. The information exchange can be divided into two phase, after a source-relay pair is selected. In the first phase, A and the selected soruce B_m simultaneously send the information to all relays, and the signal received at each relay is a superimposed signal. In the second phase, the selected relay R_n forward the received signals to two source nodes, and all other relay nodes keep idle.

Assuming a flat-fading scenario, denote the complex reciprocal channel coefficients from the nth relay R_n to A and B_m, respectively, We assume both source nodes know all channel coefficients and relay R_n only knows its own local channel coefficients h_{AR_n} and $h_{B_mR_n}$.

Consider a two-step two-way relay AF protocol with relay selection. In the first step, A and the selected source B_m transmit their signals to selected relay R_n simultaneously. Let s_A and s_{B_m} denote the symbol that will be transmitted by the source A and B_m respectively. We assume s_i is chosen from a constellation of unity power. The signal received in the nth relay can be express as

$$x_n = \sqrt{P_A} h_{AR_n} s_A + \sqrt{P_{B_m}} h_{B_mR_n} s_{B_m} + v_{R_n} \tag{1}$$

where P_{A} and P_{B_m} are the transmit powers of source node A and B_m respectively, and v_{R_n} is the complex noise at the relay R_n. In the second step, the relay R_n multiplies its received signal by a complex weight factor $\omega_{m,n}$ and broadcast the so-obtained signal to A and B_m. The signal transmitted by the relay R_n is thus $t_n = \omega_{m,n} x_n$. Let $P_R^{m,n}$ denote the transmit power of the nth relay, then the weight factor $\omega_{m,n}$ is given by

$$\omega_{m,n} = \sqrt{\frac{P_R^{m,n}}{P_A \left|h_{AR_n}\right|^2 + P_{B_m} \left|h_{B_mR_n}\right|^2 + \sigma^2}} \tag{2}$$

The signals y_A and y_{B_m} received at source node A and B_m can be represented, respectively as

$$y_A = \sqrt{P_A} \omega_{m,n} h_{AR_n}^2 s_A + \sqrt{P_{B_m}} \omega_{m,n} h_{AR_n} h_{B_mR_n} s_{B_m} + \omega_{m,n} h_{AR_n} v_{R_n} + n_A \tag{3}$$

$$y_{B_m} = \sqrt{P_{B_m}} \omega_{m,n} h_{B_mR_n}^2 s_{B_m} + \sqrt{P_A} \omega_{m,n} h_{AR_n} h_{B_mR_n} s_A + \omega_{m,n} h_{B_mR_n} v_{R_n} + n_{B_m} \tag{4}$$

where n_A and n_{B_m} is the noise at the source node A and B_m respectively. All noises are assumed to be i.i.d. Gaussian with zero-mean and unit variance. The second term on the right hand side of (3) and that of (4) are, respectively, the self-interferences of S_1 and S_2, which can be cancelled by subtracting them from the received symbols at each node. Thus, the signal at source node A and B_m after the self-interference cancellation can be written as

$$\tilde{y}_A = \sqrt{P_{B_m}} \omega_{m,n} h_{AR_n} h_{B_mR_n} s_{B_m} + \omega_{m,n} h_{AR_n} v_{R_n} + n_A \tag{5}$$

$$\tilde{y}_{B_m} = \sqrt{P_A} \omega_{m,n} h_{AR_n} h_{B_mR_n} s_A + \omega_{m,n} h_{B_mR_n} v_{R_n} + n_{B_m} \tag{6}$$

The signal $\tilde{y}_A$ and $\tilde{y}_{B_m}$ can be used to decode the information symbols s_A and s_{B_m} at source node *A* and *B*$_m$, respectively.

2.2 Problem Formulation

We consider the joint source-relay selection with power allocation for the two-way network that minimize the total transmit power with received signal to noise ratio (SNR) constraint of each node.

The main problem can thus be represented as

$$\min_{P_A, P_{B_m}, \omega_{m,n}, m,n} P_T^{m,n} \qquad \textit{subject to}: SNR_A \geq \gamma_A, SNR_{B_m} \geq \gamma_{B_m} \tag{7}$$

where γ_A and γ_{B_m} denote the SNR requirement of source node *A* and *B*$_m$ respectively.

Denote the transmit power at the *n*th relay as $P_R^{m,n}$, using (2), and assuming that the information symbols and noises are independent, it can be shown that

$$P_R^{m,n} = (P_A d_{AR_n} + P_{B_m} d_{B_m R_n} + 1)\left|\omega_{m,n}\right|^2 \tag{8}$$

where $d_{AR_n} = \left|h_{AR_n}\right|^2$ and $d_{B_m R_n} = \left|h_{B_m R_n}\right|^2$. The total transmit power, when the *i*th relay is select, $P_T^{m,n}$ can be written as

$$\begin{aligned} P_T^{m,n} &= P_A + P_{B_m} + P_R^{m,n} \\ &= P_A(1 + d_{AR_n}\left|\omega_{m,n}\right|^2) + P_{B_m}(1 + d_{B_m R_n}\left|\omega_{m,n}\right|^2) + \left|\omega_{m,n}\right|^2 \end{aligned} \tag{9}$$

The optimization problem in (7) involves continuous variables and binary variables. This is equivalent to optimizing over P_A , P_{B_m} , $\omega_{m,n}$, which is the optimal power allocation problem, then optimizing over *m*, *n*, which is the optimal source-relay selection problem. In the following, we consider the power allocation problem and source-relay selection problem separately.

3 Power Allocation and Source-Relay Selection

3.1 Power Allocation

Assuming the source node *B*$_m$ and relay node *R*$_n$ is selected, the power allocation problem can be represented as

$$\min_{P_A, P_{B_m}, \omega_{m,n}} P_T^{m,n} \qquad \textit{subject to}: SNR_A \geq \gamma_A, SNR_{B_m} \geq \gamma_{B_m} \tag{10}$$

Using (5) and (6), the receives SNRs can be written as

$$SNR_A = P_{B_m}\left|\omega_{m,n} h_{AR_n} h_{B_m R_n}\right|^2 / (1 + d_{AR_n}\left|\omega_{m,n}\right|^2) \tag{11}$$

$$SNR_{B_m} = P_A\left|\omega_{m,n} h_{AR_n} h_{B_m R_n}\right|^2 / (1 + d_{B_m R_n}\left|\omega_{m,n}\right|^2) \tag{12}$$

Obviously, the minimum total transmit power can be achieved when $SNR_A = \gamma_A$, $SNR_{B_m} = \gamma_{B_m}$. Otherwise, if, for example, the first constraint in (10) is satisfied with inequality at the optimal solution, then the optimal P_2 can be scale down to satisfy this constraint with equality. This, however, further reduces the total transmit power in (10), thereby contradicting the optimality. Thus, the transmit power of each transceiver can be represent as

$$P_A = \gamma_{B_m}(1 + d_{B_mR_n}\left|\omega_{m,n}\right|^2) / \left|\omega_{m,n}h_{AR_n}h_{B_mR_n}\right|^2 \tag{13}$$

$$P_{B_m} = \gamma_A(1 + d_{AR_n}\left|\omega_{m,n}\right|^2) / \left|\omega_{m,n}h_{AR_n}h_{B_mR_n}\right|^2 \tag{14}$$

And the equation (10) can be rewritten as

$$\min_{\omega_{m,n}} \frac{(\gamma_A + \gamma_{B_m})(1 + d_{B_mR_n}\left|\omega_{m,n}\right|^2)(1 + d_{AR_n}\left|\omega_{m,n}\right|^2)}{\left|\omega_{m,n}h_{AR_n}h_{B_mR_n}\right|^2} + \left|\omega_{m,n}\right|^2 \tag{15}$$

As can be seen from (15), the objective function does not depend on the phase of $\omega_{m,n}$. Therefore, no phase adjustment is required at the relay. Differentiating the objective function in (15) and equating it to zero lead us to the following equation:

$$-\frac{\gamma_A + \gamma_{B_m}}{\left|h_{AR_n}h_{B_mR_n}\right|^2\left|\omega_{m,n}\right|^4} + \left(\frac{\gamma_A + \gamma_{B_m}}{\left|h_{AR_n}h_{B_mR_n}\right|^2}d_{AR_n}d_{B_mR_n} + 1\right) = 0 \tag{16}$$

The positive solution to (15) given by

$$\omega^o_{m,n} = \sqrt[4]{\frac{\gamma_A + \gamma_{B_m}}{(\gamma_A + \gamma_{B_m})d_{AR_n}d_{B_mR_n} + \left|h_{AR_n}h_{B_mR_n}\right|^2}} = \sqrt[4]{\frac{\gamma_A + \gamma_{B_m}}{(\gamma_A + \gamma_{B_m} + 1)\left|h_{AR_n}h_{B_mR_n}\right|^2}} \tag{17}$$

Substituting (17) into the equation (8), then we obtain the minimum total transmit power

$$P_T^{m,n} = \frac{2}{\left|h_{AR_n}h_{B_mR_n}\right|}\sqrt{(\gamma_A + \gamma_{B_m})(\gamma_A + \gamma_{B_m} + 1)} + \left(\frac{1}{d_{AR_n}} + \frac{1}{d_{B_mR_n}}\right)(\gamma_A + \gamma_{B_m}) \tag{18}$$

Using (13) and (14), the transmit power of source node A , S_m and relay node R_n can be represent as

$$P_A = \gamma_{B_m}\left(\frac{1}{\left|h_{AR_n}h_{B_mR_n}\right|}\sqrt{\frac{\gamma_A + \gamma_{B_m} + 1}{\gamma_A + \gamma_{B_m}}} + \frac{1}{d_{AR_n}}\right) \tag{19}$$

$$P_{B_m} = \gamma_A\left(\frac{1}{\left|h_{AR_n}h_{B_mR_n}\right|}\sqrt{\frac{\gamma_A + \gamma_{B_m} + 1}{\gamma_A + \gamma_{B_m}}} + \frac{1}{d_{B_mR_n}}\right) \tag{20}$$

$$P_R^i = \frac{1}{\left|h_{AR_n}h_{B_mR_n}\right|}\sqrt{(\gamma_A + \gamma_{B_m})(\gamma_A + \gamma_{B_m} + 1)} + \frac{\gamma_A}{d_{AR_n}} + \frac{\gamma_{B_m}}{d_{B_mR_n}} \tag{21}$$

For symmetric QoS constraints where $\gamma_A = \gamma_{B_m} = \gamma$, the optimum power allocated on the relay equals the transmit power allocated on the two source node.

In high SNR regime, when $\gamma_A \gg 1$ and $\gamma_{B_m} \gg 1$, the equation (17) (18) (19) (20) and (21) can be rewritten as

$$\omega_i^o = \frac{1}{\sqrt{|h_{AR_n} h_{B_m R_n}|}} \tag{22}$$

$$P_T^{m,n} = (\gamma_A + \gamma_{B_m})(\frac{1}{|h_{AR_n}|} + \frac{1}{|h_{B_m R_n}|})^2 \tag{23}$$

$$P_R^{m,n} = (\frac{1}{|h_{AR_n}|} + \frac{1}{|h_{B_m R_n}|})(\frac{\gamma_A}{|h_{AR_n}|} + \frac{\gamma_{B_m}}{|h_{B_m R_n}|}) \tag{24}$$

$$P_A = \frac{\gamma_{B_m}}{|h_{AR_n}|}(\frac{1}{|h_{AR_n}|} + \frac{1}{|h_{B_m R_n}|}) \tag{25}$$

$$P_{B_m R_n} = \frac{\gamma_A}{|h_{B_m R_n}|}(\frac{1}{|h_{AR_n}|} + \frac{1}{|h_{B_m R_n}|}) \tag{26}$$

From (25) and (26), the transmit power of one node is proportional to the SNR requirement of the other node. When the SNR requirement of one transceiver is larger, the other transceiver will be allocated more power. The SNR requirement of each node can determine the power allocated on the selected relay and each transceiver.

3.2 Source-Relay Selection

With the optimal power allocation solution obtained, the problem in (9) reduces to the following source-relay selection problem:

$$(m^*, n^*) = \arg \min_{m \in \{1,2,\ldots,M\}} \min_{n \in \{1,2,\ldots,N\}} P_T^{m,n} \tag{27}$$

If there is only one source node, (27) can be regarded as the simplified relay selection scheme. Also, if there is only one relay node, (27) can be regarded as the simplified user selection scheme.

According to criterion (27), the source-relay selection scheme can be performed into two steps:

Step 1 is to select the best relay node for each source. The best mobile relay node R_n for each source node R_m is satisfied with

$$n' = \arg \min_{n \in \{1,2,\ldots,N\}} P_T^{m,n} \tag{28}$$

Step 2 is to select the best source-relay pair that satisfies.

$$(m^*, n^*) = \arg \min_{m \in \{1,2,\ldots,M\}} P_T^{m,n'} \tag{29}$$

During the step 1, the "best" relay can be selected either by centralized relay selection or by distributed relay selection.

According to (23), the relay selection result is dependent on the channel state information between each relay and two transceivers, and is irrelevant to the signal to noise ratio requirement of each node, γ_A and γ_{B_m} . Only the channel amplitude information is required during the source-relay selection step instead of full channel state information.

4 Performance Analysis

In this section we study the performance of joint source-relay selection and power allocation scheme in two-way opportunistic relay networks which have 6 relays, 2 sources, comparing with the power allocation according to SNR-balancing criterion in [12]. The channels coefficients are generated as zero-mean normal complex random variables. The noise power at the relays and at the two transceivers is assumed to be 0dBW. We explore the total transmit power with different SNR thresholds of two transceivers.

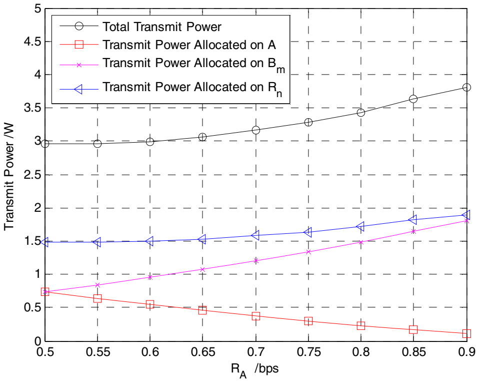


Fig. 1. Total transmit power versus R_A, for $R_A + R_{B_m} = 1$

We illustrate the transmit power required with increasing rate requirement of source node A, when the sum of rate equals to 1bps in fig.1. The power allocated on source node B_m is increasing with R_A increasing, while The power allocated on source node B_m is decreasing, as R_{B_m} decreasing. The total transmit power P_T and the power allocated on the selected relay P_{R_n} are increasing slightly with R_A increasing.

Fig.2 shows the average of the total transmit power P_T, relay transmit power P_{R_n} , and the transceiver powers P_A and P_{B_m} versus γ_A in dB. As we can see, the power allocated on A is larger than that on B_m, because the B_m need better received quality

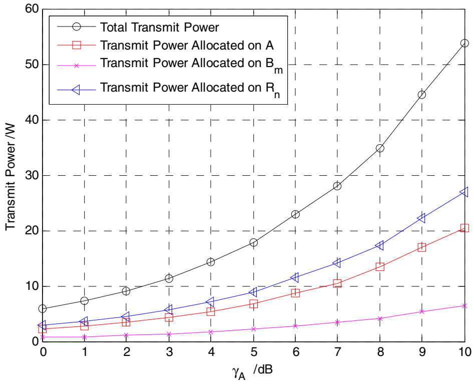


Fig. 2. Total transmit power versus γ_A, for $\gamma_{B_m} = \gamma_A + 5$

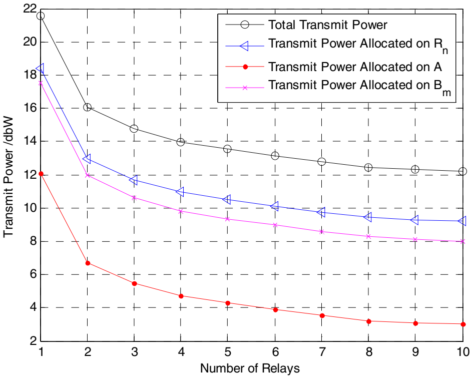


Fig. 3. Total transmit power versus number of Relays

Fig. 3 shows the minimum power required with γ_A=10dB, $\gamma_{\mathrm{B_m}} = 5\mathrm{dB}$, with different number of relays in the two way relay network. Power allocated on B_m is much larger than that on A, because the required SNR threshold of A is 5dB large than that of B_m. The power allocated on the selected relay R_n is larger than that on both transceivers, which is the same with the fig.2. With increasing number of relays, the minimum power required is decreased. Exploiting multiple relays obtains higher diversity order and lower transmit power.

In Fig.4, we compare the outage probability as a function of P_{max} for a two-way opportunistic relay system. The outage occurs when either source node falls below its threshold rate. We have set N=6, M=2, $\gamma_{\mathrm{A}} = \gamma_{\mathrm{B_m}} = 1\mathrm{dB}$ for symmetric traffics and $\gamma_{\mathrm{A}} = 1\mathrm{dB}$, $\gamma_{\mathrm{B_m}} = 5\mathrm{dB}$ for asymmetric traffics. The outage probability of two-way relay networks without power allocation is also shown. Fig.6 shows that the proposed power allocation can decrease the outage probability of TWOR-AF significantly.

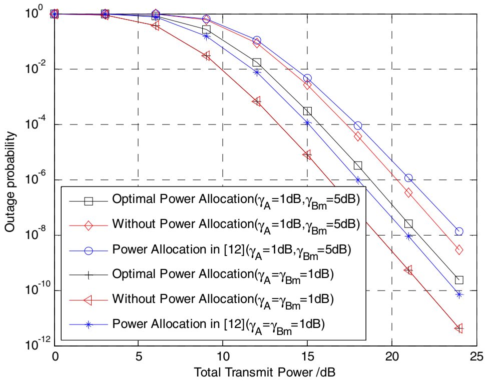


Fig. 4. Outage probability versus total transmit power, for relay number *N*=6

For symmetric traffics, the outage probability of two-way relaying with our proposed power allocation scheme equals that with the power allocation according to SNR-balancing criterion in [12]. But in asymmetric condition, the system with our proposed power allocation scheme has lower outage probability.

5 Conclusions

We developed an optimal source-relay selection scheme with power allocation for multi-source multi-relay two-way relay networks with asymmetric traffics in underground tunnels. We obtain a closed-form solution to the problem of minimizing the total transmit power subject to constraints on two transceivers' received SNRs. Then one source-relay pair which can minimize the total transmit power is selected to forward the amplified signal. We also obtained the expression of the outage probability after power location. Simulation shows that our analytic match the simulation results exactly and our proposed power allocation scheme outperforms the scheme in [12] for asymmetric traffics. Also, the outage probability is significantly decreased by using our proposed power allocation algorithm.

Acknowledgment. This work is supported by National Science Foundation of China under grant No.51274202, and the Fundamental Research Funds for the Central Universities under grant No.2013RC11.

References

1. Sun, Z., Akyildiz, I.F., Hancke, G.P.: Capacity and outage analysis of MIMO and cooperative communication systems in underground tunnels. IEEE Trans. on Wireless Communications 10(11), 3793–3803 (2011)

2. Funian, L., Desheng, Z.G.W.: Optimal power allocation for two-way relaying over OFDM using physical-layer network coding. Journal of China Universities of Posts and Telecommunications 18(1), 9–15 (2011)
3. Chen, D., Azarian, K., Laneman, J.N.: A case for amplify-forward relaying in the block-fading multiple-access channel. IEEE Transaction on Information Theory 54(8), 3728–3733 (2008)
4. Rankov, B., Wittneben, A.: Spectral efficient protocols for half-duplex fading relay channel. IEEE Journal Selected Areas Communications 25(2), 379–389 (2007)
5. Louie, R., Li, Y., Vucetic, B.: Practical physical layer network coding for two-way relay channels: Performance analysis and comparison. IEEE Transaction on Wireless Communications 9(2), 764–777 (2010)
6. Jing, Y., Jafarkhani, H.: Single and multiple-relay-selection schemes and their diversity orders. IEEE Transaction on Wireless Communications 8(3), 1414–1423 (2009)
7. Jing, Y.: A relay selection scheme for two-way amplify-and-forward relay networks. In: Proc. IEEE 2009 Wireless Communication and Signal Processing, Nanjing, China, November 13-15, pp. 1–5 (2009)
8. Song, L.: Relay Selection for two-way relay with amplify-and-forward protocols. IEEE Transaction on Vehicular Technology 60(4), 1954–1959 (2011)
9. Zhang, X., Gong, Y.: Adaptive power allocation in two-way amplify-and-forward relay networks. In: Proc. IEEE 2009 International Conference Communications (ICC 2009), Dresden, Germany, June 14-18, vol. 4, pp. 1–5 (2009)
10. Liu, T.C.-K., Xu, W., Dong, X., Lu, W.-S.: Adaptive power allocation for bidirectional amplify-and-forward multiple-relay multiple user networks. In: Proc. IEEE Globe Telecommunications Conference (GLOBECOM 2010), Miami, America, December 6-10, pp. 1–6 (2010)
11. Han, Y., Ting, S., Ho, C., Chin, W.: High-rate two-way amplify-and-forward half-duplex relaying with OSTBC. In: Proc. IEEE Vehicular Technology Conference (VTC 2008), Marina Bay, Singapore, May 11-16, pp. 2426–2430 (2008)
12. Zhang, Y., Ma, Y., Tafazolli, R.: Power allocation for bidirectional AF relaying over Rayleigh fading channels. IEEE Communication Letter 14(2), 145–147 (2010)
13. Talwar, S., Jing, Y., Shahbazpanahi, S.: Joint relay selection and power allocation for two-way relay networks. IEEE Signal Process Letter 18, 91–94 (2011)

Performance Analysis of Broadcast in Multi-channel Multi-radio Wireless Mesh Networks

Min Song and Xiaohua Xu

The University of Toledo, Toledo, OH 43606, USA

Abstract. Broadcast is a fundamental operation for wireless mesh networks. It plays an important role in the communication protocol design. Many existing work have studied the NP-hard broadcast problem in multi-hop networks. However, most of them assume a single-channel and single-radio wireless network model. We investigate broadcast in multi-channel multi-radio wireless mesh networks. In multi-channel multi-radio wireless mesh networks, the wireless interference due to simultaneous transmissions from the same channel and intra-node interference render the broadcast problem nontrivial. In this work, we analyze the performance of a broadcast protocol with MAC-layer scheduling under different networking conditions. We also explore the performance improvement by incoporating the neighbor elimination scheme with the broadcast protocol. We analyze the performance improvement of the integrated protocol in environments of multi-channel multi-radio and multi-rate.

Keywords: Wireless mesh networks, broadcast, multi-channel multi-radio, performance analysis.

1 Introduction

Wireless mesh networks have received great attentions due to their various applications, such as broadband home networking, community and neighborhood networking, and enterprize networking. Several cities and wireless companies around the world have deployed mesh networks. The United States Armed Forces (USAF) are using wireless mesh networks to connect their computers, mainly laptops, in field operations as well. In this application, the mesh networks can enable troops to know the locations and conditions of every soldier or marine, and to coordinate activities without much direction from central command. Mesh networks have also been used as the last mile solution for extending the Internet connectivity for mobile wireless nodes. Wireless mesh networks consist of two types of nodes: mesh routers and mesh clients [1]. Mesh routers form an infrastructure for mesh clients. A small set of routers also function as gateways connecting to the wired network. Single-radio single-channel mesh networks suffer from a serious capacity degradation [12]. To improve the network capacity, a promising approach is to provide each node with multi-channel multi-radio (MC-MR) capabilities and to permit MAC protocols to adjust the transmission rate [7]. In MC-MR wireless mesh networks, the radios or antennas at each node are independent with each other, and the channels are also orthogonal to each other.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 110–120, 2013.

In multi-hop MC-MR wireless mesh networks, network-wide broadcast from a distinguished source node to other nodes is a primitive communication task. For example, exchanging control information among neighboring nodes, such as channel availability and routing information, is crucial for the realization of most networking protocols in ad-hoc networks. The data control information are often sent out as broadcasts. In addition, some exigent data such as emergency messages and alarm signals are also delivered as broadcasts. Thus, broadcast becomes a key research topic in network community and has been extensive studied recently in different scenarios [3, 8, 9, 14, 15, 21–24].

We will consider a MC-MR mesh network model which expands the available spectrum and reduces interference. In MC-MR wireless mesh networks, a set of communications can occur successfully at the same time as long as (1) each radio is involved in at most one communication, and (2) over each channel, all communications are conflict-free under the single-radio single-channel setting. Message passing on different channels result in different groups of neighborhood. Broadcast in MC-MR wireless mesh networks is known to enjoy the *wireless broadcast advantage* [5, 18]. In wireless networks, as long as an omni-directional antenna is used, the transmission power corresponds to the coverage range in all directions. One transmission is enough to deliver a message to all devices within the range. In addition, the presence of multi-radio allows a mesh node to send and receive at the same time; the availability of multi-channel allows channels to be reused across the network for broadcasting.

In this work, we conduct performance analysis of broadcast under different networking conditions. The performance metrics contain reliability, goodput, delay, redundancy, overhead, and scalability. We will adopt a comprehensive link and channel quality metrics which can quantify the data rate, reliability, and so on. The broadcast protocol assigns channels and radios for message passing procedure while reduces wireless interference at the same time. We propose to determine the threshold for selecting good links and channels. Addtitionally, we propose a MAC-layer scheduling heuristic which relies on spatial reuse to maximize the throughput. We also explore the performance improvement by employing the neighbor elimination scheme in the broadcast protocol in environments of MC-MR and multi-rate. In the neighbor elimination scheme, each node that receives the message for the first time does not broadcast it immediately, but waits for a given duration, which can be computed or randomly generated. We analyze the performance improvement of the integrated protocols in the scenario of MC-MR and multi-rate.

The rest of the paper is organized as follows. Section 2 provides the problem formulation. Section 3 provides the details of performance analysis. Section 4 briefly surveys the related work. The conclusions and future work are given in Section 5.

2 Problem Statement

In a MC-MR wireless mesh network, each node has at least one radio. Each radio is tuned to one of the available non-overlapping channels. We can use a directed graph $G = (V, E)$ to model the MC-MR wireless mesh network. Here V is the set of nodes and E is the set of communication links. Each link is represented by an ordered quintuple containing the transmitter node, the receiver node, the radio at the transmitter node, the radio at the receiver node, and the transmitting channel. For simplicity, we also use $\overrightarrow{uv}$ to denote a communication link from the node u to v. We assume that the nodes are placed with a constraint of connectivity.

Two types of wireless interference are considered. The first one is the inter-node interference, which occurs when adjacent nodes are using the same channel. We assume the protocol interference model [12] and both the communication range and interference ranges are uniform. The second type of wireless interference is the intra-node interference, which occurs when multiple channels are used by the same node. In MC-MR mesh networks, when a node beacons a packet, not all of its one-hop neighbors can receive this packet. A node can receive a packet from a neighboring transmitter node only if the node is tuned to the same wireless channel; otherwise, it cannot receive the packet. The collision may happen on different channels concurrently. This is different from single-channel single-radio network where data communication occurs only on a global common channel.

There is a distinguished source node which broadcasts messages to all the other nodes in the wireless mesh network. The objective is to analyze the performance of the broadcast protocol [21] under different networking conditions. The analytical metrics contain reliability, goodput, broadcasting delay, broadcasting redundancy, overhead, and scalability. Note that we will consider the tradeoff among different performance metrics. If there is one metric such as minimize the number of transmissions, the problem can be easily formulated as an integer Linear Programming (LP) model [24]. The LP model has been used for the problems of multi-commmodity flow, minimum power broadcast, maximum lifetime broadcast, and so on.

Another objective is to explore the performance improvement by employing the neighbor elimination scheme with the broadcast protocol in the MC-MR and multi-rate scenario. That is, each node that receives the message for the first time does not broadcast it immediately, but waits for a given duration, which can be computed or randomly generated. In Figure 1, there are three nodes s_1, s_2, s_3 that receive a message for the first time. Suppose there is only one channel for transmitting. If they broadcast the message immediately, there will be conflict at the receiver node v_3, v_4, v_5. If we employ the neighbor elimination scheme, s_1 broadcasts in the first time-slot, s_3 broadcasts in second time-slot. Then all the receiver nodes can receive the message without conflict. In addition, the node s_2 does not need to broadcast.

3 Performance Analysis

In this section, we analyze the performance of the channel model and the broadcast protocol with MAC-layer scheduling under different networking conditions.

3.1 Preliminaries

We briefly summarize the channel model and the broadcast protocol proposed in [21].

In the channel model, the *link quality* of a communication link $\overrightarrow{ij}$ with channel c is quantified as $w_{ij,c} = R_c \cdot DR_{ij,c}$. Here R_c is the transmission rate of channel c, and $DR_{ij,c}$ is the packet delivery rate [6, 7] of the communication link $\overrightarrow{ij}$ with channel c.

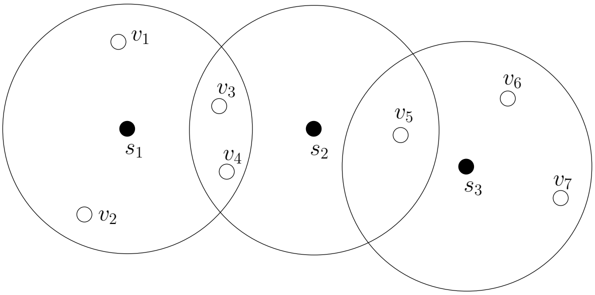


Fig. 1. The illustration of the neighbor elimination scheme

For each node i, the *channel quality* of a channel c depends on the qualities of all incident links that use the channel and is defined as

$$w_{i,c} = R_c \cdot \sum_{j \in \mathbf{N}_c(i)} DR_{ij,c}/|\mathbf{N}(i)|$$

Here $N(i)$ is the set of nodes within the communication range of node i and $N_c(i)$ is the set of nodes within the communication range of node i and tuned to channel c for receiving.

Based on the channel model, the broadcast protocol consists of two phases. The first phase is constructing local structures. Each node figures out good transmission channels and good incident links. The second phase is building the broadcasting tree. By using a control message passing procedure, each node uses its local structure and the ones from its neighbors to build a local broadcasting branch. The local broadcasting branch fully exploits the channel diversity and assigns channels and radios for broadcasting to reduce the wireless interference. The union of these branches results in a broadcasting tree. The second phase can be represented by a finite state machine.

When constructing the quasi-optimal broadcast protocol, a common channel is used to exchanging all the control messages. The broadcast protocol can ensure that all nodes in the network receive the broadcasting messages effectively and efficiently. Additionally, the broadcast protocol satisfies the following principles:

1. Each node has only one receiving channel;
2. There is no need for a node to participate in broadcasting if all of its neighbors are already covered;
3. For each message, each node may broadcast more than once by using different channels on different radios;
4. Each node should not use the same channel for both transmission and reception;
5. Adjacent nodes should avoid using the same channel for transmissions.

3.2 Determining Thresholds for Channel and Link Selection

Recall that the first phase of the broadcast protocol is removing all bad channels and links whose weights are below the thresholds. After the first phase, we focus on broadcast by only using the remaining channels and links instead. In other words, only the

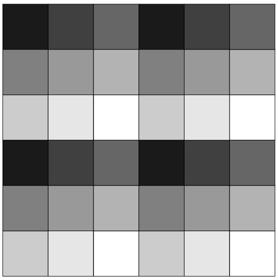


Fig. 2. The illustration of the partition and coloring scheme

good links and channels of weights at least the link threshold and channel threshold respectively will participate in broadcasting. Therefore, determining the thresholds becomes a critical issue and has never been investigated before.

The thresholds depend on many factors, such as the network environment, and network traffic. In reality, different wireless networks may have different thresholds. We propose a binary-search-based heuristic to find the channel threshold and link threshold.

Start with initial values big enough for thesholds, each time we check whether the current values of the channel threshold and link threshold can ensure that the directed network G is strongly connected. Here, a directed path in which only good links are included is called a Strongly Connected Path (SCP); a directed network is strongly connected if there is at least one SCP between any pair of nodes.

If the current values for the channel threshold and link threshold cannot ensure the strong connectivity of the network, we decrease both values by half and repeat the checking procedure. We terminate until the channel threshold and link threshold can ensure the strong connectivity of the network.

We make a note that there must exist values for the channel threshold and link threshold which can ensure the strong connectivity of the network. The reason is that the nodes are placed with a constraint of connectivity. When the values of the channel threshold and link threshold decrease, the number of good links and channels increase, When the values of the channel threshold and link threshold are small enough, there exists a SCP between any pair of nodes, *i.e.*, the network is strongly connected.

3.3 MAC-Layer Scheduling

We propose the following MAC-layer scheduling heuristic. The heuristic relies on the spatial reuse and schedules as many links as possible each time.

Let us partition the deployment region into cells and color the cells. The partition and coloring scheme can ensure that when at most one node for every grid of a monotone color transmits, the transmissions are interference-free.

Based on the partition and coloring scheme (Figure 2), each time we select a color in a round-robin manner. For each cell of the selected color, we find all nodes lying inside this cell which contain data to transmit; we choose the node with the smallest level in the broadcast tree and find the corresponding link (from this node to its children) in the broadcast tree. Note that, one link may connect to multiple nodes. All the found links form an independent set which can transmit concurrently. We then try to insert links greedily subject to the wireless interference constraint; We repeat the above procedure until all data have been broadcasted to each part of the network.

If the number of channels and radios is very large, we allow the fine-grained MAC-layer scheduling heuristic to degenerate into a simple greedy algorithm.

3.4 Comprehensive Analytical Model

The performance analysis is conducted in terms of reliability, goodput, delay, redundancy, overhead, and scalability. For comparative analysis, the pure flooding algorithm is considered, in which a channel is chosen to distribute messages to every part of the network.

Reliability Metric. The reliability of a message is the percentage of the number of nodes that has received this message. The reliability of the system is the average reliability of all messages.

To prove full reliability, we need to prove that the broadcasting tree obtained from the broadcast protocol is strongly connected. The argument is based on the concept of SCP. A directed broadcasting tree is strongly connected if there is at least one SCP from the distinguished source node to any other node. We can verify that as long as the original network is strongly connected, the broadcast protocol outputs a strongly connected broadcasting tree with its depth upper-bounded. Thus, the broadcast protocol consistently achieves full reliability.

On the other hand, the pure flooding algorithm cannot achieve full reliability due to serious contentions and wireless interference. Some broadcast messages may be dropped. At the same time, the wireless interference among nearby nodes causes collisions.

When the message disseminating speed increases, the MAC-layer scheduling can ensure that all messages are broadcasted to each part of the network without collisions. Thus, the broadcast protocol with the MAC-layer scheduling scheme can achieve full reliability. On the other hand, the pure flooding algorithm has to handle the new broadcasting messages while the previous ones still stay in the buffer. Thus, the broadcast messages keep being accumulated in each node's buffer. The broadcasting of the new messages is interleaved with that of the accumulated messages. To sum up, collisions happen not only in the same broadcast message, but also among the consecutive messages.

Even without the neighbor elimination scheme, the broadcast protocol can achieve full reliability as the broadcasting tree is strongly connected. After integrating the neighbor elimination scheme, the pure flooding algorithm's reliability increases. The reason is that the neighbor elimination scheme can resolve the heavy contention problem.

Goodput Metric. The goodput of a node is the number of useful information bits delivered by the source node to this node, per unit of times. The goodput of the system is the average goodput of all nodes across the network. For simplicity, we can assume the packet length is fixed.

Each non-redundant received message contributes to the goodput. The goodput of a node far away from the source node is small since the delivery of useful messages costs a lot of time. When the number of nodes increases, the average distance from the source node increases accordingly; the average goodputs of the broadcast protocol and the pure flooding algorithm decrease. There exists a limit for the goodput of the system when the number of nodes increases.

After integrating the neighbor elimination scheme, the broadcast protocol's goodput decreases since the lapesed time for delivering a message from the source node to each node increases. The pure flooding algorithm's goodput increases as the neighbor elimination scheme can resolve the heavy contention problem and help the useful information delivery.

Delay Metric. The worst case delay of a message is the lapsed time from the time when the source node starts beaconning the message to the time when the last node receives this message. The average worst case delay is the average worst case delay of all messages.

The broadcast protocol can ensure interference-freeness. When the number of nodes increases, the average worst case delay of the broadcast protocol is only increased slightly. The reason is that the length of the longest path from the source node is increased sub-linearly. On the other hand, the pure flooding algorithm has a large wost case delay due to receiving redundancies, serious collisions, and wireless interference. When the number of nodes increases, the average worst case delays of both the broadcast protocol and the pure flooding algorithm increase. Generally, the average worst case delay varies inversely with the goodput.

After integrating the neighbor elimination scheme, we conjecture that the average worst case delay of the broadcast protocol increases while the average worst case delay of the pure flooding algorithm decreases.

Redundancy Metric. If any message is received by a node for more than once, this message is perceived as a duplicate message for this node. The receiving redundancy of a node is the total number of duplicate messages across all messages. The average receiving redundancy is the average of the receiving redundancy of each node across the network.

The broadcast protocol significantly reduces the receiving redundancy. This is because only the nodes included in the broadcast tree relay the broadcast messages and only the nodes that tune to the same channel as the transmitter nodes receive the broadcast messages. On the other hand, the number of neighboring nodes increases as the number of nodes increases. Thus, the average receiving redundancy of the pure flooding algorithm increases linearly as the number of nodes increases.

The neighbor elimination scheme can reduce the number of receiving neighbors each time. Thus, the neighbor elimination scheme can reduce the average receiving redundancy promisingly.

The Transmission Redundancy: The transmission redundancy of a message is the percentage of the number of nodes participating in broadcasting this message. The average transmission redundancy is the average of the transmission redundancy over all messages.

The transmission redundancy of the broadcast protocol only depends on the percentage of non-leaf nodes in the broadcasting tree. The pure flooding algorithm has a transmission redundancy close to one since every node participats in broadcasting messages if the message is not dropped. When the number of nodes is increased, the transmission redundancy of the broadcast protocol remains the same approximately while the transmission redundancy of the pure flooding algorithm decreases. This is due to the fact that the reliability of the pure flooding algorithm decreases, and thus fewer nodes participate in the broadcasting.

After integrating the neighbor elimination scheme, the transmission redundancy of the broadcast protocol remains the same. The transmission redundancy of the pure flooding algorithm increases as the reliability increases.

Message Overhead. The message complexity is defined as $\sum_{i=1}^{N}\sum_{j=1}^{M} X_{i,j}$, where $X_{i,j}$ is the total number of the j-th packet (including duplicates) received by node i. The message complexity is related to the receiving redundancy metric. The broadcast protocol significantly reduces the message complexity.

The number of control messages to construct the broadcast tree increases linearly with the total number of communication links.

Scalability. The scalability of the system is the ability to handle a growing amount of communication traffics in a capable manner or its ability to be enlarged to accommodate that growth. When constructing the broadcast tree, at most two-hop information is needed. Thus, the distributed broadcast protocol has good scalability. On the other hand, when the message disseminating speed becomes high, the network traffic becomes high. If there are too many nodes within a small region, this region becomes a bottleneck in terms of the network traffic; traffic congestions may occur in this region. Thus, the pure flooding algorithm cannot scale well.

After integrating the neighbor elimination scheme, the broadcast protocol still has good scalability if the neighbor elimination scheme is scalable. In the neighbor elimination scheme, only local information is needed, thus the neighbor elimination scheme has good scalability.

4 Related Work

4.1 Wireless Broadcasting

The classical broadcast problem in wireless networks have been extensively studied. Chlamtac and Kutten [4] proved the NP-hardness of finding a minimum makespan collision-free broadcast scheduling for general graphs, even in the absence of wireless interference. Gandhi *et al.* [9] proved the NP-hardness of the minimum makespan broadcast scheduling problem for multihop wireless networks with bounded transmission range and proposed a simple 468-approximation algorithm. Huang *et al.* [14] improved the approximation ratio to be 16.

Two widely used broadcasting methods are the probabilistic and tree-based approaches. In the probabilistic broadcasting approach (also called gossip-based approach) [2, 10, 11, 13, 17], when a node first receives a broadcasting message, it broadcasts the message to its neighbors with a probability of p and drops the message with a probability of $1 - p$. The challenges of the probabilistic approach are how to find the appropriate gossiping parameters and how to guarantee full reliability. In the tree-based approach [10], a broadcasting tree is constructed first to reduce redundant transmissions before the broadcasting messages are actually transmitted. The tree-based method involves an overhead to construct the tree.

Most of the broadcasting protocols have been developed primarily with one focus: reliability, broadcast delay, or redundant transmissions. These performance metrics are often contradictory goals. In an effort to minimize delay and the number of retransmissions, a broadcast schedule is developed for collision free broadcasting [10]. While the results are promising, the assumption of a single-radio single-channel and single-rate model limits its usage in MC-MR networks. [20] presented a set of algorithms to achieve low broadcasting delay in MC-MR and multi-rate mesh networks. The broadcasting tree is constructed using a set of centralized algorithms with a goal of minimizing broadcasting delay. However, the centralized approach results in a nontrivial overhead to construct and maintain the tree. In addition, these algorithms are evaluated in a 10-node mesh network, thus making it less clear about the scalability of the proposed algorithms.

The most related work to this paper is [21] where a novel channel model is developed to assess the channel quality in the presence of interfering networks, and distributed broadcasting protocols are developed. In-depth theoretical analysis has been conducted and a simulator has been developed to simulate MC-MR and multi-rate ad hoc mesh networks. Recently, [22] proposed a Quality-of-Service (QoS)-based broadcast protocol for multi-hop cognitive radio ad-hoc networks under blind information. In [23], a distributed broadcast protocol is proposed in multi-hop cognitive radio ad-hoc networks without a common control channel.

4.2 Channel Quality Assessment

Interference-aware routing protocols [7] and interference-aware MAC layer protocols [16] usually assume that either a priori information about the interference is known, or a 0-1 function is applied to the link, *i.e.*, a link either works or does not work. Few studies have contributed to defining the measurement of interference. The first exceptional study was made in [19] to estimate the link interference in a static single-radio single-channel experimental wireless network. The way this study calculates the interference is not practical in real-world mesh networks. Thus, finding a practical wireless interference-aware metric is crucial. [6] presented a metric termed as Expected Transmission Count (ETX) to find a high throughput path. The ETX of a link is calculated using the forward and reverse delivery rates of the link. The ETX of a path is then the sum of the ETX for each link in the path. Although ETX does very well in homogeneous single-radio environments, it does not perform well in environments with multiple radios as indicated in [7]. We further argue that ETX does not accurately represent the quality of the entire path in the context of broadcasting as no acknowledgment exists in broadcasting.

5 Conclusion

We have conducted the performance analysis of broadcast in MC-MR wireless mesh networks. For the future work, we will conduct real-world experiments. Additionally, we will explore cognitive radio mesh network broadcasting as cognitive radio network is emerging as a promising paradigm for the future networks to relieve the spectrum scarcity and hence signicantly improve spectrum utilization.

Acknowledgement. The authors would like to thank Changlong Chen, Tao Zou and Zhuang Zhang for their participations in discussions about this research.

The research of Min Song is supported in part by NSF grant CNS-1248092 and NSF IPA Independent Research and Development (IR/D) Program. However, any opinion, finding, and conclusions or recommendations expressed in this material; are those of the author and do not necessarily reflect the views of the National Science Foundation. The research of Xiaohua Xu is supported in part by NSF grant CNS-1248092.

References

1. Akyildiz, I.F., Wang, X., Wang, W.: Wireless mesh networks: a survey. Computer Networks 47(4), 445–487 (2005)
2. Boyd, S., Ghosh, A., Prabhakar, B., Shah, D.: Gossip algorithms: Design, analysis and applications. In: IEEE INFOCOM, pp. 1653–1664 (2005)
3. Chen, Z., Qiao, C., Xu, J., Lee, T., et al.: A constant approximation algorithm for interference aware broadcast in wireless networks. In: IEEE INFOCOM, pp. 740–748 (2007)
4. Chlamtac, I., Kutten, S.: On broadcasting in radio networks–problem analysis and protocol design. IEEE Transactions on Communications 33(12), 1240–1246 (2002)
5. Cui, T., Chen, L., Ho, T.: Distributed optimization in wireless networks using broadcast advantage. In: IEEE Conference on Decision and Control, pp. 5839–5844 (2007)
6. De Couto, D.S., Aguayo, D., Bicket, J., Morris, R.: A high-throughput path metric for multihop wireless routing. Wireless Networks 11(4), 419–434 (2005)
7. Draves, R., Padhye, J., Zill, B.: Routing in multi-radio, multi-hop wireless mesh networks. In: ACM MOBICOM, pp. 114–128 (2004)
8. Gandhi, R., Kim, Y.-A., Lee, S., Ryu, J., Wan, P.-J.: Approximation algorithms for data broadcast in wireless networks. IEEE Transactions on Mobile Computing 11(7), 1237–1248 (2012)
9. Gandhi, R., Mishra, A., Parthasarathy, S.: Minimizing broadcast latency and redundancy in ad hoc networks. IEEE/ACM Transactions on Networking 16(4), 840–851 (2008)
10. Gandhi, R., Parthasarathy, S., Mishra, A.: Minimizing broadcast latency and redundancy in ad hoc networks. In: ACM MOBIHOC, pp. 222–232 (2003)
11. Gavidia, D., Voulgaris, S., Van Steen, M., et al.: A gossip-based distributed news service for wireless mesh networks. In: Annual Conference on Wireless On-demand Network Systems and Services, pp. 59–67 (2006)
12. Gupta, P., Kumar, P.R.: The capacity of wireless networks. IEEE Transactions on Information Theory 46(2), 388–404 (2000)
13. Haas, Z.J., Halpern, J.Y., Li, L.: Gossip-based ad hoc routing. In: IEEE INFOCOM, pp. 1707–1716 (2002)
14. Huang, S., Wan, P., Jia, X., Du, H., Shang, W.: Minimum-latency broadcast scheduling in wireless ad hoc networks. In: IEEE INFOCOM, pp. 733–739 (2007)

15. Huang, S.-H., Wu, H.-C., Iyengar, S.S.: Multisource broadcast in wireless networks. IEEE Transactions on Parallel and Distributed Systems 23(10), 1908–1914 (2012)
16. Kyasanur, P., Vaidya, N.H.: Routing and link-layer protocols for multi-channel multi-interface ad hoc wireless networks. ACM SIGMOBILE Mobile Computing and Communications Review 10(1), 31–43 (2006)
17. Li, X.-Y., Moaveninejad, K., Frieder, O.: Mobile Networks and Applications. Regional gossip routing for wireless ad hoc networks 10(1-2), 61–77 (2005)
18. Neely, M.J., Urgaonkar, R.: Opportunism, backpressure, and stochastic optimization with the wireless broadcast advantage. In: IEEE Asilomar Conference on Signals, Systems and Computers, pp. 2152–2158 (2008)
19. Padhye, J., Agarwal, S., Padmanabhan, V.N., Qiu, L., Rao, A., Zill, B.: Estimation of link interference in static multi-hop wireless networks. In: USENIX IMC, pp. 28–28 (2005)
20. Qadir, J., Misra, A., Chou, C.T.: Minimum latency broadcasting in multi-radio multi-channel multi-rate wireless meshes. In: IEEE SECON, pp. 80–89 (2006)
21. Song, M., Wang, J., Xing, K., Park, E.: Interference-aware broadcasting in multi-radio multi-channel mesh networks. IEEE Transactions on Wireless Communications 7(12), 5473–5481 (2008)
22. Song, Y., Xie, J.: A qos-based broadcast protocol for multi-hop cognitive radio ad hoc networks under blind information. In: IEEE GLOBECOM, pp. 1–5 (2011)
23. Song, Y., Xie, J.: A distributed broadcast protocol in multi-hop cognitive radio ad hoc networks without a common control channel. In: IEEE INFOCOM, pp. 2273–2281 (2012)
24. Wang, J., Song, M., Hsieh, G., Xin, C.: Minimum cost broadcast in multi-radio multi-channel wireless mesh networks. In: International Conference on Mobile Ad-hoc and Sensor Networks, pp. 238–247 (2011)

SAFE: A Strategy-Proof Auction Mechanism for Multi-radio, Multi-channel Spectrum Allocation*

Ruihao Zhu, Fan Wu**, and Guihai Chen

Shanghai Key Laboratory of Scalable Computing and Systems
Shanghai Jiao Tong University, China
zhurh1992@sjtu.edu.cn, {fwu,gchen}@cs.sjtu.edu.cn

Abstract. The rapid growth of wireless technology has led to increasing demand for spectrum. In the past, spectrum is statically allocated. As a result, many wireless applications cannot use idle spectrum even though it is left unused by the owner for a long period of time. The low utilization of already scarce spectrum resource requires us to dynamically reallocate the idle spectrum to achieve better spectrum usage. In this paper, we model the problem of spectrum reallocation as a sealed-bid reserve auction, and propose SAFE, which is a Strategy-proof Auction mechanism For multi-radio, multi-channel spEctrum allocation. We prove the strategy-proofness of SAFE theoretically, and evaluate its performance extensively. Evaluation results show that SAFE achieve good performance, in terms of spectrum utilization and buyer satisfaction ratio.

1 Introduction

Radio spectrum is under great demand due to the latest development of wireless technology. The static allocation of radio spectrum can no longer meet the growth of wireless applications since it not only leaves lots of radio spectrum unused in some geographic areas, but also makes the idle radio spectrum unavailable to new wireless applications that do not have licensed spectrum bands. Consequently, dynamic radio spectrum reallocation is highly needed to solve or alleviate the problem of spectrum shortage.

A feasible way to reallocate radio spectrum is to perform auction. The Federal Communications Commission (FCC) has adopted this method for approximately two decades [3]. In contrast to FCC, which only holds auctions for large buyers, we concentrate on small buyers like private wireless networks and mobile wireless applications.

However, there are two major difficulties in designing efficient mechanisms for radio spectrum auction. One difficulty, coming from the auction theory itself, is strategy-proofness (see Section 3.2 for the definition of strategy-proofness). Generally speaking,

* This work was supported in part by the State Key Development Program for Basic Research of China (Grant No. 2012CB316201), in part by China NSF grant 61272443, 61133006, 61073152, and in part by Shanghai Science and Technology fund 12PJ1404900 and 12ZR1414900. The opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies or the government.

** Corresponding author.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 121–134, 2013.

strategy-proofness means that any buyer cannot get higher payoff by bidding a value other than her true valuation for the goods. In the radio spectrum auction, the buyers are rational and always try their best to maximize their own payoffs. The buyers may manipulate the auction to seek for more benefit, and thus lower the spectrum utilization and hurt the other buyers' interests. As a result, designing a strategy-proof mechanism for the spectrum auction is of undoubted importance. The other difficulty comes from the reusability of the radio spectrum. The reusability of the radio spectrum can allow two buyers to use the same spectrum simultaneously as long as they have enough distance (out of the interference range) between each other geographically. What's more, the optimal solution for the problem of spectrum allocation is NP-complete [1, 19] in multi-hop wireless networks.

In the past, several works on strategy-proof auction mechanisms for spectrum allocation have been proposed [12], For example, Zhou et al.'s VERITAS [20] and TRUST [21], Wu and Vaidya's SMALL [15]. VERITAS uses the method of critical neighbor to solve a single side radio spectrum auction. However, it does not take the valuation of the sellers into account and the seller's utility might be negative. TRUST uses McAfee mechanism to perform a double side auction and it could guarantee the profit of the sellers, but it has to waste a channel and sacrifice a lot of good buyers. SMALL introduces the concept of reserve price to guarantee the profit of seller, and would only sacrifice a bounded number of buyers. However, when applied to multi-radio, multi-channel spectrum auction, SMALL can only work under a very strong assumption that the channels must be sold out.

The existing problems of the previous mechanisms prompt us to design an auction mechanism which can allocate spectrum efficiently in general situations and protect the interests of the seller. In this paper, we present a Strategy-proof Auction mechanism For multi-radio, multi-channel spEctrum allocation (SAFE). In SAFE, there is a seller with multiple radio spectrum channels and has a reserve price for each of the channel. All buyers submit their sealed bids at the beginning of the auction so that they do not have any information about the others' bids. The channel can be reused and each buyer can get more than one channel. The auction mechanism selects the winning buyers and allocates channels to them iteratively. The seller may not sell a channel if the bid for this channel is lower than the reserve price. The major advantages of SAFE are listed as follow:

- SAFE enables a multi-radio, multi-channel spectrum auction, in which all the buyers can bid for more than one channel and get more than one channel as well.
- When compared with VERITAS, SAFE has higher spectrum utilization and buyer satisfaction ratio, which are two commonly used metrics for evaluating channel auction mechanisms. Spectrum utilization is the sum of the number of channels that the buyers get divided by the number of channels; and the buyer satisfaction ratio is the proportion of the buyers who gets at least a given ratio of their requested channels.
- When compared with TRUST, the seller can possibly sell all the channels; while in TRUST, the seller always has to waste one channel. SAFE sacrifices much less buyers than TRUST, and the number of sacrificed buyers by SAFE is bounded by the number of channels.

- When compared with SMALL, SAFE does not depend on the very strong assumption that the channels must be sold out.

The major contributions in this paper are as follows: We propose a strategy-proof auction mechanism for multi-radio, multi-channel spectrum auction, namely SAFE, and prove its strategy-proofness. We also implement SAFE, and numerically evaluate its performance. We list our detailed contributions as follows:

- **Strategy-Proofness:** We prove that SAFE can achieve both incentive-compatibility and individual rationality. When SAFE is applied, the dominant strategy of each buyer is to bid her true valuation, and none of them is charged higher than her true valuation for a channel when bidding truthfully.
- **Seller Incentives:** SAFE adopts the concept of reserve price, and thus guarantees that the seller is profitable by selling her channels.
- **Extensive Evaluation:** We conduct extensive simulations to evaluate SAFE's performance. The results verify that SAFE does guarantee strategy-proofness, and achieves good performance in terms of spectrum utilization and buyer satisfaction ratio.

The rest of this paper is organized as follows. In Section 2, we review related works. In Section 3, we present technical preliminaries. In Section 4, we introduce the multi-radio, multi-channel spectrum auction mechanism— SAFE, and prove that it is a strategy-proof mechanism. In Section 5, we give the evaluation results of SAFE. In Section 6, we draw conclusions and point out possible future work directions.

2 Related Works

In this section, we briefly review existing works on radio spectrum auctions.

Zhou et al. proposed VERITAS [20], which uses critical price to determine the charges for the channel winners. Although VERITAS can be applied to multi-radio, multi-channel auction, it suffers from low spectrum utilization due to the nature of the greedy channel allocation. Later, they proposed TRUST [21], which is an elegant double side mechanism adopting the thoughts from McAfee mechanism. Wu and Vaidya have proposed SMALL [15], which incorporates the concept of reserve price to protect the benefit of the channel seller, and achieves superior performance to TRUST. Although SMALL is the closest work to ours, it cannot be applied to general multi-radio, multi-channel auction, because it relies on the assumption that the channels are scarce compared with the users. In contrast to the above works, SAFE can guarantee strategy-proofness for multi-radio, multi-channel auction in the general case, and achieve high spectrum utilization and buyer satisfaction ratio.

Online spectrum auction is also an important topic of spectrum allocation. Hajiaghayi et al. [8] studied the supply-limited online auction model and proposed a value- and time-strategy-proof mechanism, which can achieve constant efficiency and revenue-competitiveness. Hajiaghayi et al. [7] also provided a characterization for truthful online auction rules. Later, Xu et al. [17] proposed an efficient online spectrum auction mechanism that can decide whether to permit each user's exclusive usage of spectrum and calculate their corresponding charges. Deek et al. [2] presented Topaz that can allocate spectrum efficiently and achieve strategy-proofness.

There are also some existing game theory based works on spectrum allocation that are not based on auction [6, 18]. Felegyhazi et al. [4] studied Nash equilibria in a static multi-radio, multi-channel allocation game, and proposed two algorithms to converge the system to the Nash equilibria. Later, Wu et al. [16] presented a stronger mechanism for the above problem to make the system converge to a stable state in a single step.

3 Technical Preliminaries

In this section, we present our auction model for multi-radio, multi-channel spectrum allocation, and briefly review some important solution concepts from mechanism design.

3.1 Auction Model of Spectrum Allocation

We consider a secondary spectrum market. There is a "seller", who is a wireless communication infrastructure provider. She has a number of channels, and wants to sell her idle channels to wireless service providers who do not have official licences on radio spectrum from the government. The wireless service providers are called the "buyers", and want to buy the licenses of the idle channels from the seller to provide services to their customers. We model the problem as a sealed-bid reserve auction. The seller sets a reserve price for each channel for sale. The reserve price of a channel can be regarded as the maintaining cost of the channel. The buyers submit their sealed bids simultaneously, such that the buyers cannot know each other's bid. If the bid (or the group bid from a group of buyers) for a channel is less than the channel's reserve price, the seller can refuse to sell this channel.

In the auction, the seller can acts as the auctioneer, if she is trustworthy; otherwise, a trusted central authority is required to perform as the auctioneer. The seller has a set of orthogonal channels for sale, denoted by $\mathbb{C} = \{c_1, c_2, \ldots, c_m\}$. As we mentioned, the seller has a reserve price s_k for each channel $c_k \in \mathbb{C}$. We denote the profile of reserve prices by

$$\boldsymbol{s} = (s_1, s_2, \ldots, s_m).$$

Each channel can be used by more than one buyers, if they are out of the interference range of each other.

We denote the set of buyers by $\mathbb{N} = \{1, 2, \ldots, n\}$. We assume that the buyers do not have preference over the channels. Each buyer $i \in \mathbb{N}$ has a per-channel valuation v_i, which is private information of the buyer. This is commonly known as *type* in the literatures. The valuation for a channel can be the revenue gained from providing wireless services using the channel. In the auction, the buyers choose their bids, denoted by

$$\boldsymbol{b} = (b_1, b_2, \ldots, b_n),$$

which are based on their types, and simultaneously submit the sealed bids to the auctioneer. If a buyer $i \in \mathbb{N}$ is equipped with multiple radio interfaces, she can also claim to request up to r_i channels. In contrast to the per-channel valuation, we assume that the buyers do not cheat about the maximal number of channels they want. This assumption is based on the fact that the buyers do not have the incentives to cheat the maximal number of requested channels. On one hand, if a buyer requests more than enough channels, she may need to pay for extra channels that are not needed at all. On the other hand, if a buyer under claims the number, she definitely cannot win the expected number of channels in the auction. We denote the profile of channel requests by

$$\boldsymbol{r} = (r_1, r_2, \dots, r_n).$$

The buyers also submit their channel requests together with the bids to the auctioneer.

The auction mechanism determines the set of winning buyers, channel allocation, and charges for the winners. Let's denote the set of channels won by buyer $i \in \mathbb{N}$ by $A_i \subseteq \mathbb{C}$, and the charge for buyer i on each channel $c_k \in A_i$ by p_i^k. We can denote the channel allocation by

$$\boldsymbol{A} = (A_1, A_2, \dots, A_n),$$

where $\forall i \in \mathbb{N}, 0 \leq |A_i| \leq r_i$. We now can define the utility u_i of each buyer $i \in \mathbb{N}$ to be the difference between her total valuation and total charge on the channels won:

$$u_i = v_i|A_i| - \sum_{k \in A_i} p_i^k.$$

Clearly, if a buyer does not win any channel in the auction, then both her charge and utility are zero.

We assume that the buyers are rational, and their objectives are to maximize their own utilities. We also assume that the auction is collusion-free.

In contrast, the objective of the auction mechanism is to increase spectrum utilization and buyer satisfaction ratio. Here, spectrum utilization is the sum of channels the winning buyers get diveded by the number of channels $\frac{\sum_{k=1}^{n}|A_k|}{m}$; buyer satisfaction ratio is the proportion of buyers who get at least a given ratio of their requested channels in the auction.

3.2 Solution Concepts

In this section, we introduce the solution concepts used in this paper from mechanism design.

A strong solution concept from mechanism design is *dominant strategy*.

Definition 1 (Dominant Strategy [5,10]). *Strategy a_i is a player i's dominant strategy, if for any $a_i' \neq a_i$ and any strategy profile of the other players a_{-i},*

$$u_i(a_i, a_{-i}) \geq u_i(a_i', a_{-i}).$$

Intuitively, a dominant strategy of a player is one that maximizes her utility, regardless of what strategies the other players choose. In our sealed-bid reserve auction for spectrum allocation, the strategy of a buyer $i \in \mathbb{N}$ is her bid $b_i = a_i(v_i)$.

The concept of dominant strategy is the basis of *incentive-compatibility*, which means that there is no incentive for any player to lie about her private information, and thus revealing truthful information is the dominant strategy for every player. A company concept is *individual-rationality*, which means that for every player who faithfully participate the game/auction is expected to gain no less utility than staying outside. We now can introduce the definition of *Strategy-Proof Mechanism*.

Definition 2 (Strategy-Proof Mechanism [9] [11]). *A mechanism is strategy-proof when it satisfies both incentive-compatibility and individual-rationality.*

4 SAFE

In this section, we present the design of SAFE, and prove its strategy-proofness.

4.1 Design of SAFE

When designing SAFE, we follow the design rational of SMALL, but with significant differences. SAFE contains three algorithms: buyer grouping, winner selection, and charge determination.

(1) Buyer Grouping
SAFE first constructs a conflict graph of the buyers. In the conflict graph, each node represents a buyer. Any pair of buyers who lie within the interference range of each other are connected by an edge in the conflict graph. SAFE then groups the buyers using an existing graph coloring algorithm (*e.g.*, [14]), so that no buyer can be assigned to more than one group and the buyers who are connected are not in the same group. Since the grouping algorithm is bid-independent, no buyer can manipulate this process. We denote the buyer groups by

$$\mathbb{G} = \{g_1, g_2, \ldots, g_q\},$$

where

$$g_j \cap g_l = \emptyset, \forall g_j, g_l \in \mathbb{G},$$

$$\text{and} \quad \bigcup_{g_j \in \mathbb{G}} g_j = \mathbb{N}.$$

We regard each group as a super buyer, and define a group bid for each group $g_j \in G$ as:

$$B_j = (|g_j| - 1) \cdot \min\{b_i | i \in g_j\}.$$

If more than one buyer report the smallest bid in the group, we regard the one with the smallest alphabetical order as the smallest-bid bidder. We denote the smallest-bid buyer in group g_j by $SMALLEST(g_j)$. We now get a profile of group bids:

$$\boldsymbol{B} = (B_1, B_2, \ldots, B_q).$$

(2) Winner Determination

SAFE determines the winners and channel allocation iteratively. In each iteration, SAFE sorts the remaining channels by the reserve price in a non-decreasing order, and sorts the remaining buyer groups by the group bids in a non-increasing order:

$$\mathcal{C}^t : c_1^t, c_2^t, \ldots, c_{m^t}^t, \quad s.t., s_1^t \le s_2^t \le \ldots \le s_{m^t}^t,$$

$$\mathcal{G}^t : g_1^t, g_2^t, \ldots, g_q^t, \quad s.t., B_1^t \ge B_2^t \ge \ldots \ge B_q^t.$$

Here, t indicates that this is in the tth iteration, and m^t denotes the current number of remaining channels. Since the number of groups does not change in different iterations, we can use the same q in all the iterations. If two channels have the same reserve price or two groups have the same group bid, the order between them is determined following the alphabetical order.

Next, SAFE finds the maximal possible number of trades l^t in the tth iteration, *s.t.*

$$\sum_{i=1}^{l^t} s_i^t \le \sum_{i=1}^{l^t} B_i^t. \tag{1}$$

Finally, SAFE selects the first l^t groups in list $\mathcal{G}^t$ as winning buyer groups, and assigns the first l^t channels in list $\mathcal{C}^t$ to the corresponding winner groups. In each winning group, the buyer(s), except the one who bids the smallest in the group, are winning buyers and can get the channel assigned to the group. We denote the set of winning buyers in the tth iteration by

$$W^t = \bigcup_{j=1}^{l^t} \{i | i \in g_j^t \wedge i \ne SMALLEST\left(g_j^t\right)\}.$$

For each winning buyer $i \in W^t$, SAFE decreases its number of requested channels r_i by 1. If $r_i = 0$, then buyer i's demand is fully filled, and SAFE removes buyer i from the buyer group g_j she belongs to and updates group g_j's group bid:

$$g_j = g_j \setminus \{i\},$$

$$B_j = (|g_j| - 1) \cdot \min\{b_i | i \in g_j\}.$$

SAFE also deletes the channels, which have already been sold, from the set of channels.

SAFE repeats the above procedure until no more winner can be generated (*i.e.*, $l^t = 0$).

The pseudo-code of above winner selection and channel allocation is shown by Algorithm 1. In Algorithm 1, function $GROUPING(\mathbb{N})$ is a graph coloring based grouping algorithm, and returns the buyer grouping result.

We note that in each trade, only one buyer, who bids the smallest in her buyer group, is sacrificed. As a result, the number of buyers sacrificed does not exceed m, which is the number of channels for sale.

Let d denote the largest degree in the conflict graph of the buyers. The computational complexity of the greedy graph coloring based buyer grouping algorithm is $O(n + |E|)$, and the number of groups is at most $(d + 1)$. In each iteration, SAFE takes

Algorithm 1. Winner Determination and Channel Allocation Algorithm

```
Require: A set of channels $\mathbb{C}$, a profile of reserve prices $\boldsymbol{s}$, the number of channels $m$, a set of
    buyers $\mathbb{N}$, a profile of bids $\boldsymbol{b}$, and a profile of channel requests $\boldsymbol{r}$.
Ensure: A set of winning buyers $\mathbb{W}$ and a profile of channel allocation $\boldsymbol{A}$.
$\mathbb{W} \leftarrow \varnothing$; $\boldsymbol{A} \leftarrow \varnothing^n$; $t \leftarrow 1$; $m^t \leftarrow m$.
$(\mathbb{G}, q) \leftarrow GROUPING(\mathbb{N})$.
repeat
    for all $g_j \in \mathbb{G}$ do
        $B_j = (|g_j| - 1) \cdot min\{b_i | i \in g_j\}$.
    end for
    Sort the channels $\mathbb{C}$ by reserve price $\boldsymbol{s}$ in non-decreasing order: $c_1^t, c_2^t, \ldots, c_{m^t}^t$, s.t., $s_1^t \leq$
    $s_2^t \leq \ldots \leq s_{m^t}^t$.
    Sort buyer groups $\mathbb{G}$ by group bid $\boldsymbol{B}$ in non-increasing order: $g_1^t, g_2^t, \ldots, g_q^t$, s.t., $B_1^t \geq$
    $B_2^t \geq \ldots \geq B_q^t$.
    $l^t \leftarrow \underset{l^t \leq \min\{m^t, q\}}{argmax} \left( \sum_{i=1}^{l^t} s_i^t \leq \sum_{i=1}^{l^t} B_i^t \right)$.
    $W^t \leftarrow \bigcup_{j=1}^{l^t} \{ i | i \in g_j^t \wedge i \neq SMALLEST\left(g_j^t\right) \}$.
    $\mathbb{W} \leftarrow \mathbb{W} \cup W^t$.
    for $j \leftarrow 1$ to $l^t$ do
        for all $i \in g_j^t \setminus \{SMALLEST\left(g_j^t\right)\}$ do
            $A_i \leftarrow A_i \cup \{c_j^t\}$; $r_i \leftarrow r_i - 1$.
            if $r_i = 0$ then
                $g_j^t \leftarrow g_j^t \setminus \{i\}$.
            end if
        end for
        $\mathbb{C} \leftarrow \mathbb{C} \setminus \{c_j^t\}$.
    end for
    $m^{t+1} \leftarrow m^t - l^t$; $t \leftarrow t + 1$.
until $l^t = 0$.
return $\mathbb{W}$ and $\boldsymbol{A}$.
```

$\max\{O(m \log m, O(d \log d))\}$ time to sort the bids and reserve price and $O(m)$ time to determine the number of good trades. What's more, SAFE can run at most m iterations. Therefore, the overall computational complexity of SAFE is $O(n + |E| + m \cdot \max\{m \log m, d \log d\})$.

(3) Charging

For each channel $k \in A_i$ a buyer $i \in g_j$ wins, the charge p_i^k is equal to the smallest bid in g_j. So the total charge for a buyer $i \in g_j$ is:

$$p_i = \sum_{k \in A_i} p_i^k = |A_i| \cdot \min\{b_k | k \in g_j\}.$$

We note that the charge for each buyer is independent of her bid.

The seller receives all the charges from the winning buyers. We assume that Algorithm 1 iterates for η times. Then the total revenue of the seller is:

$$\begin{aligned} Revenue &= \sum_{t=1}^{\eta}\sum_{j=1}^{l_t}(|g_j^t|-1)\cdot min\{b_i|i\in g_j^t\} \\ &= \sum_{t=1}^{\eta}\sum_{j=1}^{l_t} B_j^t. \end{aligned} \tag{2}$$

From Inequation (1) and Equation (2), we get that the seller's profit is always non-negative in the auction:

$$\begin{aligned} Profit &= Revenue - \sum_{t=1}^{\eta}\sum_{j=1}^{l^t} s_j^t \\ &= \sum_{t=1}^{\eta}\sum_{j=1}^{l_t} B_j^t - \sum_{t=1}^{\eta}\sum_{j=1}^{l^t} s_j^t \\ &\geq 0. \end{aligned}$$

4.2 Strategy-Proofness

In order to show that SAFE is a strategy-proof auction mechanism for multi-radio, multi-channel spectrum allocation, we prove the following lemma first.

Lemma 1. *If SAFE is used, each buyer's dominant strategy is reporting her per-channel valuation as a bid.*

Proof. We prove this by showing that by bidding untruthfully, any buyer cannot get a higher utility in any iteration. We consider a buyer $i \in g_j$ with per-channel valuation v_i. Let $b_j^{min} = \min\{b_i|i \in g_j\}$. Since the smallest-bidding buyer in group g_j cannot win any channel, b_j^{min} remains the same in all the iterations. Consequently, buyer i's utility u_i^t in the tth iteration is either $v_i - b_j^{min}$ or 0, depending on whether she wins a channel or not in this iteration. Let $\hat{u}_i^t$ be the utility got by buyer i in the tth iteration, when she bids truthfully. We next prove that buyer i cannot increase her utility got in any iteration, by distinguishing two cases:

1. Buyer i is the smallest-bidding buyer in group g_j, when bidding truthfully (*i.e.*, $b_i = v_i$). Then, her utility $\hat{u}_i^t = 0, \forall t$. We further distinguish two cases:
   - If she increases her bid to $b_i' > v_i$, then the smallest bid in group g_j becomes $b_j'^{min} \geq b_j^{min} = v_i$. If she still does not win a channel, her utility remains 0:

     $$u_i'^t = 0 = \hat{u}_i^t.$$

     If she wins a channel in the tth iteration, then her utility for this iteration becomes non-positive:

     $$u_i'^t = v_i - b_j'^{min} \leq v_i - v_i = 0.$$

   - If she decreases her bid to $b_i' < v_i$, then she is still the smallest-bidding buyer in group g_j, and her utility remains 0.
2. Buyer i is not the smallest-bidding buyer in group g_j, when bidding truthfully. We further distinguish three cases:
   - If she increases her bid to $b_i' > v_i$, then the smallest bid in group g_j remains unchanged. Consequently, her utility also remains unchanged, no matter she wins a channel or not in this iteration.
   - If she decreases her bid to $b_i' < v_i$ and becomes the smallest-bidding buyer in the group g_j, then she definitely cannot win a channel and her utility is $u_i'^t = 0 \leq \hat{u}_i^t$.
   - If she decreases her bid to $b_i' < v_i$ but is still not the smallest-bidding buyer in the group g_j, then the group bid of g_j remains unchanged. No matter the group g_j is a winning group in the tth iteration or not, the buyer i's utility remains unchanged:

$$\begin{cases} u_i'^t = v_i - b_j^{min} = \hat{u}_i^t, & g_j \text{ is a winning group;} \\ u_i'^t = 0 = \hat{u}_i^t, & \text{otherwise.} \end{cases}$$

Since buyer i maximizes her utility gain in each iteration by bidding truthfully (*i.e.*, $b_i = v_i$), and the total utility of buyer i is the sum of her utilities gained in all the iterations, we can conclude that buyer i maximizes her utility by bidding truthfully. This completes our proof.

Lemma 2. *SAFE guarantees the individual rationality.*

Proof. By the charging scheme, for each channel won, a buyer $i \in g_j$ is not charged more than her bid. If the buyer i bids truthfully, then the per-channel charge is also no larger than her per-channel valuation:

$$\begin{aligned} \hat{u}_i &= v_i|A_i| - \sum_{k \in A_i} p_i^k \\ &= |A_i|(v_i - \min\{b_k | k \in g_j\}) \\ &\geq 0. \end{aligned}$$

Since SAFE satisfies both incentive compatibility and individual rationality, we have the following theorem.

Theorem 1. *SAFE is a strategy-proof auction mechanism multi-radio, multi-channel spectrum allocation.*

5 Numerical Results

We implement SAFE and evaluate its performance in terms of buyer utility, spectrum utilization and buyer satisfaction ratio.

5.1 Methodology

We run SAFE for over 1000 times to evaluate its performance. The terrain area is 1800 meters × 1800 meters, and the buyers are randomly distributed in this area. The number of buyers varies from 50 to 800 with step of 50. The number of channels is 12 or 24, and each buyer can request up to 5 channels. The interference range of the buyers is 425 meters. The reserve prices of the channels lie in the range of $(0, 1]$, and the buyers' per-channel valuations also lie in the same range $(0, 1]$.[1] We group the buyers using a greedy graph coloring algorithm [13].

5.2 Metrics

We use three metrics to measure SAFE's performance:

- *Buyer Utility*: For each buyer, her utility is the difference between her total valuation and total charge on the channels won. We distinguish two kinds of buyer behaviors that may result in different buyer utilities.
  - *Bidding Truthfully*: Bidding truthfully means that the bid submitted by a buyer is her true per-channel valuation.
  - *Misreporting*: Misreporting means that a buyer submits a bid other than her true per-channel valuation. In the evaluation, we assume that a misreported bid also lies in $(0, 1]$.
- *Spectrum Utilization*: Spectrum utilization is the sum of the number of channels that the buyers get divided by the number of channels.
- *Buyer Satisfaction Ratio*: Buyer satisfaction ratio is defined as the proportion of the buyers who get at least a given ratio of their requested channels.

Since TRUST and SMALL cannot guarantee strategy-proofness for multi-radio, multi-channel spectrum auction in general cases, we only compare the performance of SAFE with VERITAS.

5.3 Buyer Utility

We numerically verify the strategy-proofness of SAFE with 500 buyers and 24 channels. We randomly choose a buyer and investigate her utilities with different behaviors. We run SAFE for over 1000 times. In each run, we fix the other buyers' bids, and evaluate the chosen buyer's utilities of truthfully bidding and misreporting. After that, we randomly choose 50 records to show. We note that the rest records lead to the same conclusion.

Fig. 1 shows the utilities of buyer 178 with different node behaviors. According to the figure, we can find that buyer 178's utility of bidding truthfully is always at least as high as that of misreporting in each run. Furthermore, her utilities of bidding truthfully are always nonnegative, while misreporting can even lead to negative utilities (*e.g.*, utility of misreporting is negative in the 20th, 33rd and 46th run).

[1] The ranges of buyers' per-channel valuations and seller's reserve prices can be chosen differently from the ones used here. However, the evaluation results of using different ranges are similar to each other. As a result, we only show the results for the above ranges in this paper.

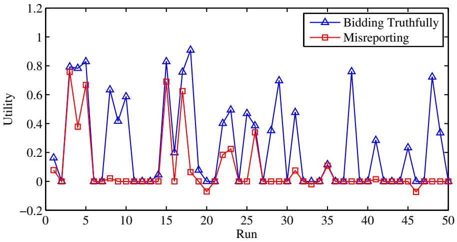
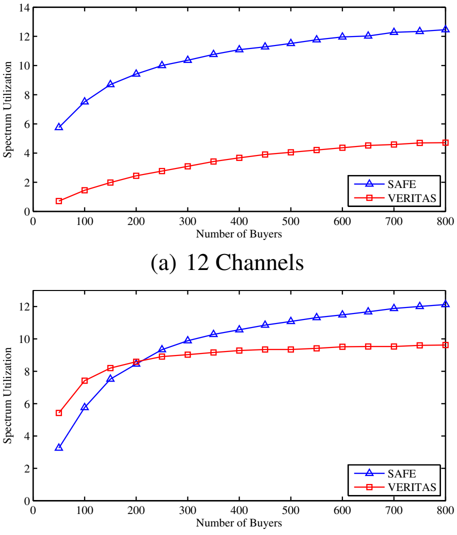


Fig. 1. Utilities of buyer 178 when bidding truthfully and misreporting

5.4 Spectrum Utilization

In this subsection, we show the comparison results on spectrum utilization between SAFE and VERITAS.



(a) 12 Channels

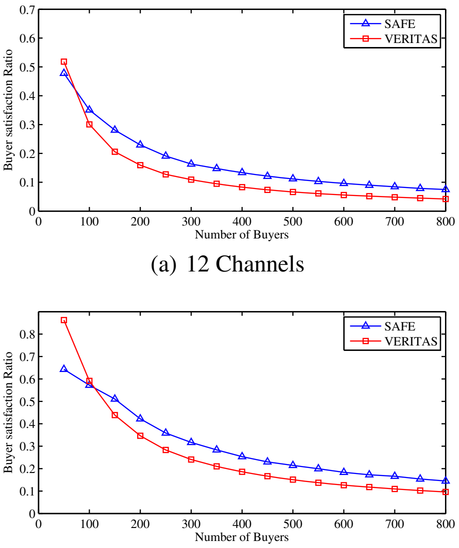
(b) 24 Channels

Fig. 2. Spectrum utilization of SAFE and VERITAS

Fig. 2(a) and Fig. 2(b) show the comparison results on spectrum utilization of SAFE and VERITAS with 12 and 24 channels, respectively. The results show that SAFE outperforms VERITAS in terms of spectrum utilization in most of the situations. The only exception is when there are 24 channels and the number of buyers is small (*i.e.*, less than 200). This is because when the total number of channels requested by the buyers is relatively small compared with the number of channels, VERITAS can possibly fill large proportion of the buyers' request. The results also prove that SAFE is suitable for the situations in which spectrum is a scarce resource.

5.5 Buyer Satisfaction Ratio

Finally, we compare the buyer satisfaction ratios of SAFE and VERITAS. Here, for fair comparison, we set the threshold ratio to be 0.5, which means that a buyer in SAFE is satisfied if she gets at least 50% of her requested channels.



(a) 12 Channels

(b) 24 Channels

Fig. 3. Buyer satisfaction ratio of SAFE and VERITAS

Fig. 3(a) and Fig. 3(b) show the comparison results on buyer satisfaction ratios of SAFE and VERITAS with 12 and 24 channels, respectively. From the figures, the results show that SAFE outperforms VERITAS in terms of buyer satisfaction ratio in most of the cases. The only exception is when the number of buyers is small (*i.e.*, less than 100 buyers for 12 channels and less than 150 buyers for 24 channels), VERITAS has a higher buyer satisfaction ratio. This is because SAFE has to sacrifice the buyers who bid the least in the buyer groups to guarantee strategy-proofness.

From the above results, we can draw the conclusion that SAFE is an efficient strategy-proof auction mechanism for multi-radio, multi-channel spectrum allocation.

6 Conclusions and Future Work

In this paper, we have modeled the problem of multi-radio, multi-channel spectrum allocation as a sealed-bid auction. We have proposed SAFE, which achieves both strategy-proofness and high system performance, in terms of spectrum utilization and buyer satisfaction radio. As for future work, we are going to design collusion-resistant auction mechanisms for the multi-radio, multi-channel spectrum allocation.

References

1. Cox, D.C., Reudink, D.O.: Dynamic channel assignment in high capacity mobile communication system. Bell System Technical Journal 50(6), 1833–1857 (1971)
2. Deek, L., Zhou, X., Almeroth, K., Zheng, H.: To preempt or not: Tackling bid and time-based cheating in online spectrum auctions. In: INFOCOM 2011 (April 2011)
3. Federal Communications Commission (FCC), `http://www.fcc.gov/`
4. Félegyházi, M., Čagalj, M., Bidokhti, S.S., Hubaux, J.-P.: Non-cooperative multi-radio channel allocation in wireless networks. In: INFOCOM 2007 (May 2007)
5. Fudenberg, D., Tirole, J.: Game Theory. MIT Press (1991)
6. Gao, L., Wang, X.: A game approach for multi-channel allocation in multi-hop wireless networks. In: MobiHoc 2008 (December 2008)
7. Hajiaghayi, M.T.: Online auctions with re-usable goods. In: EC 2005 (2005)
8. Hajiaghayi, M.T., Kleinberg, R., Parkes, D.C.: Adaptive limited-supply online auctions. In: EC 2004 (2004)
9. Mas-Colell, A., Whinston, M.D., Green, J.R.: Microeconomic Theory. Oxford Press (1995)
10. Osborne, M.J., Rubenstein, A.: A Course in Game Theory. MIT Press (1994)
11. Varian, H.: Economic mechanism design for computerized agents. In: USENIX Workshop on Electronic Commerce (1995)
12. Wang, X., Li, Z., Xu, P., Xu, Y., Gao, X., Chen, H.: Spectrum sharing in cognitive radio networks – an auction based approach. IEEE Transactions on System, Man and Cybernetics–Part B: Cybernetics 40, 587–596 (2010)
13. Welsh, D.J.A., Powell, M.B.: An upper bound for the chromatic number of a graph and its application to timetabling problems. The Computer Journal 10(1), 85–86 (1967)
14. West, D.B.: Introduction to Graph Theory, 2nd edn. Prentice Hall (1996)
15. Wu, F., Vaidya, N.: SMALL: A strategy-proof mechanism for radio spectrum allocation. In: INFOCOM 2011 (April 2011)
16. Wu, F., Zhong, S., Qiao, C.: Globally optimal channel assignment for non-cooperative wireless networks. In: INFOCOM 2008 (April 2008)
17. Xu, P., Xu, X., Tang, S., Li, X.-Y.: Truthful online spectrum allocation and auction in multichannel wireless networks. In: INFOCOM 2011 (April 2011)
18. Yu, Q., Chen, J., Fan, Y., Shen, X.S., Sun, Y.: Multi-channel assignment in wireless sensor networks: A game theoretic approach. In: INFOCOM 2010 (April 2010)
19. Yue, W.: Analytical methods to calculate the performance of a cellular mobile radio communication system with hybrid channel assignment. IEEE Transactions on Vehicular Technology 40(2), 453–460 (1991)
20. Zhou, X., Gandhi, S., Suri, S., Zheng, H.: eBay in the sky: Strategy-proof wireless spectrum auctions. In: MobiCom 2008 (September 2008)
21. Zhou, X., Zheng, H.: TRUST: A general framework for truthful double spectrum auctions. In: INFOCOM 2009 (April 2009)

A Context-Aware MAC Protocol for VANETs

Qiu Xu, Haikuan Fan, Shuangfei Guo, and Xiaoqiang Xiao

Computer School,
National University of Defense Technology,
Changsha, China

Abstract. Collision-free transmission is important for reliable and delay-bounded wireless communication, which is required by many safety-related applications in vehicular ad-hoc networks (VANETs). However, with the increase of the vehicle node density, the channel contention collision of IEEE 802.11p known as the media access control (MAC) protocol standard for VANETs increases as well. Aiming at this problem, we propose a context-aware MAC protocol for VANETs with the basic idea of ensuring only one vehicle node access the channel initiatively while others conceal their contend intentions. According to the context message, we use the hamming competing network to decide which node will access the channel. Simulation results show that the proposed protocol has a considerably low collision probability and high transmission reliability while keeping a low access delay even in high dense scenario.

Keywords: VANETs, 802.11p, MAC, hamming network, fuzzy logic.

1 Introduction

The immense number of fatal accidents and traffic jams resulted from the increasing number of vehicles on road has driven the research and development of new-generation technologies that help drivers travel more safely and efficiently. In recognition to these problems, the IEEE community is working on the standardization of IEEE 802.11p[1], which intends to enhance the IEEE 802.11 to support vehicular ad hoc networks (VANETs) applications where reliability and low delay are crucial.

At the same time, in Europe, ETSI integrated a slightly modied version of IEEE 802.11p in a European Standard for Intelligent Transportation Systems (ITS)[2]. Although it is still an object of debate whether to use multiple radios or periodic channel switching, the general agreement is that most safety-related applications have selected $100ms$ to be an interval of synchronization of which vehicles will synchronize to the CCH first and then alternate to the SCH during the remaining time.

The safety-related applications on CCH are based on broadcasting periodically (single-hop) Cooperative Awareness Messages (CAM) or so-called beacons which include data such as current position, speed, and acceleration. Based on this knowledge, the drivers can better operate vehicles to avoid potential dangers or jams since

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 135–150, 2013.

every vehicle is aware of other vehicles within a certain range. To achieve this goal, all vehicles should have fair access to CCH to disseminate its beacons or some other safety-related messages. However, the CSMA/CA mechanism adopted by 802.11p protocol couldnt handle the fairness problems properly in many scenarios .Where some nodes occupy the channel consistently while others starve[3]. This problem tends to be more obvious when the density of vehicles is high[4].

To solve this problemwe try to reduce the access competition conflict with the beacon message. From the beacon message, we can get the position, velocity and acceleration information which provide a basis for the design of the MAC mechanism.

In the proposed protocol each contending node will get a set of priorities of all the contending nodes based on the context information and the one itself with the highest priority will gain the opportunity to access the channel, adaptive back-off mechanism is introduced to solve the potential collisions caused by the nonideal channel or emergency in the complex vehicle environment. Therefore, only one node will access the channel initiatively without other contending nodes and achieves the goal of being collision-free.

The major contributions of our paper are summarized as follows:

a. Redefine the beacon message format, provides a basis to determine the vehicle node access priority.

b. Processing position, speed, acceleration information of vehicle with fuzzy logic method.

c. Take hamming competing networks to determine the access node.

d. Designing the adaptive back-off mechanism to resolve potential active competition problem and wait for the deadlock problem.

The rest of this paper is organized as follows. Section 2 presents a review of the significant contributions in the scope of VANETs MAC protocols found in the literature. The characterization of our proposed protocol and its algorithm are introduced in Section 3. In Section 4, we analyze the proposed MAC protocol in terms of collision probability, time delay and reliability based on simulation results. Section 5 concludes this paper and point out the future work.

2 Related Work

Most of vehicular safety applications proposed in the literature rely on the IEEE 802.11p standard.Therefore, recent studies have focused on improvement of CSMA/CA which contributes to the poor performance of 802.11p such as unfair access, unavoidable collisions, unpredictable access delay and consecutive packet drops, especially in dense scenario. These problems are mainly caused by the fixed initial CW (contention window) from which a random back-off time was selected to avoid collisions while the increasing number of nodes makes the matters worse. To the best of our knowledge the problem of dynamically adapting the CW size for reliable broadcast in VANETs was first discussed in [5], but receivers may not hear the short safety messages sent towards them so that the estimation of traffic load condition may not be accurate.

In addition to the above random access protocols, deterministic protocols have also been studied extensively since they are near collision-free and guarantee a finite access time to the channel, even in a dense scenario. The most popular deterministic method is time division multiple access (TDMA)[6], where the timeline is split into a series of the time periods, and each period is divided into a set of time slots. Each car is then assigned a slot in which it transmits its messages every period. It usually requires a centralized controller and this represents a big impediment in the case of vehicle-to-vehicle (V2V) communications. A distributed TDMA-based near collision-free MAC protocol named Reservation Protocol (RP)[7] proposes a proactive approach called pre-scheduling, in the sense that nodes gather information, and schedule future time slots with their prospective neighbors in advance. Including protocols proposed in [8] [9], they usually need knowledge of neighboring topology which will introduce a lot of overhead and communication delay, and they cannot differentiate the priorities of emergency messages and common applications.

3 Context-Aware MAC Protocol

Our proposed context-aware MAC protocol aims at making a dense and dynamic network seem to be relatively sparse and stable, reducing the contention collisions, packet drops while ensuring low latency and fairness.We define a nodes context information as the global status information of all the nodes that share the same channel with it, i.e., all the broadcasted beacons shared among them.

The flow graph of our protocol is illustrated in Fig.1 and the context-awareness is realized by three steps: priority prediction, hamming competition and adaptive back-off. Fig.2 shows the status of the proposed protocol during an interval of synchronization. When a node receives a beacon from a neighbor node (the time between t_0 and t_1, it indicates that the channel is occupied by this neighbor node for the current time and the context information get updated. At time t_2 when the neighbor releases the channel after a $100ms$ occupation, collisions will be

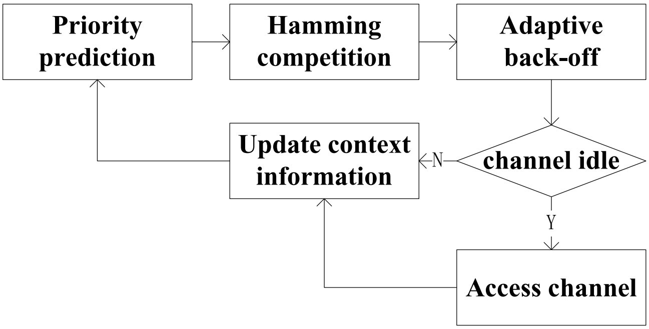


Fig. 1. Proposed protocol flow graph

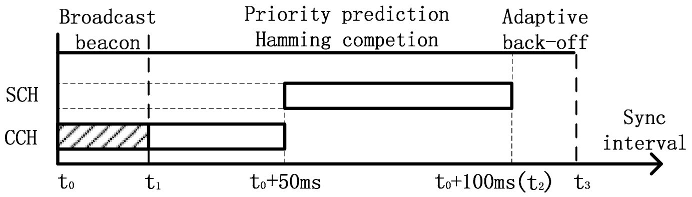


Fig. 2. Status during an interval of synchronization

unavoidable if other nodes access the channel with CSMA/CA. During the time between receiving beacon from a neighbor node and the time the neighbor node releases the channel (time t_1 to t_2), all the other nodes waiting for accessing the channel will take the following two steps respectively: (1) predict priorities of all the nodes based on the updated context information, (2) select the one with the highest priority. Generally speaking, since all the nodes involved in the contention implement step (1) based on the same global context information, each node in step (2) will select a common competing winner which will access the channel by itself while the losers will wait for the context updating and take the above two steps again.However, the winner in step (2) may face potential contentions from nodes that cannot obtain exact context information due to nonideal channel or those with emergencies. Therefore, step (3) is needed to make the node not access the idle channel until the back-off time expires (at time t_3). The back-off time is selected randomly from a CW which is also context-aware.After a collision-free channel access and successful transmission, the context information get updated and a new round channel contention begins with the above three steps again.

3.1 Priority Prediction

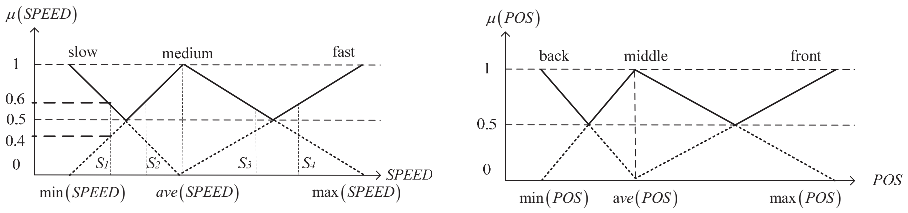
As mentioned above, the context based on which to predict priorities of all the contending nodes is actually a set of historical beacons broadcasted among them. Therefore, we predict each nodes priority based on its own broadcasted beacon including: TS(Timestamp of the beacon), $TYPE$(Type of the node), POS(GPS information of the node), $SPEED$(Speed of the node),ACC(Acceleration of the node). We use $i(TS, TYPE, POS, SPEED, ACC)$ to represent the interested fields where the subscript i stands for ID which is the unique identification of the node.In this part we predict four sub-priorities according to each field respectively and combine them into an integrated competing priority $w_i(TS, TYPE, POS, SPEED, ACC)$ in part B. We assume that the subscript $i = 1, 2 \ldots M$, where M is the number of contending nodes, i.e., the number of IDs.

TS Priority $w_i(TS)$. Since the wireless channel cannot be shared by several nodes at the same time, the TS fields of beacons are different from each other. We define $\Delta t_i = t_{current} - b_i(TS)$ which means the gap between the last time node i access the channel and the current time. The bigger this gap is, the hungrier the node is. We sort the nodes in descending order of gap. The one with the biggest gap obtains the highest TS priority $(00)_2$ while the smallest gap obtains the lowest TS priority $(M-1)_2$. The number of bits in the binary TS priority is $\lceil log_2 M \rceil$, which is decided by M, the number of contending nodes.

TYPE Priority $\boldsymbol{w_i(TYPE)}$**.** The emergency public transportation includes police car, ambulance, fire engines, etc. Obtain the highest priority $(00)_2$. Common public transportation such as taxi, bus gets the second priority $(01)_2$ and then followed by large cars like trucks and coaches$(10)_2$. The lowest priority is assigned to private car $(11)_2$.

SPEED and POS Priority $\boldsymbol{w_i(SPEED, POS)}$**.** The contending nodes that share the same channel and can exchange beacon messages discussed above are all locate in a one-hop range, called communication range in this paper. Research results in [10] and [11] indicate that compared with the average speed of the nodes in the communication rage, the one with higher relative speed tend to have less time to stay in the communication range, therefore, it should have a higher priority to ensure it access the channel as soon as possible before it leaves the communication range. But they didn't consider the relative position and the acceleration of each node in the communication range, for example, a accelerated node moves faster than the average speed but lies in the back (compared to the moving direction)will still have a lot of time to pass through the communication range. Therefore, it doesn't need to have a higher priority even with a higher relative speed. Only the one which moves faster and faster lies in the front should have higher priorities to access the channel since they are moving out of the communication range. We take this into consideration here to combine the relative speed and the relative position to improve the priority prediction algorithm.

The concept of fast speed or back position is difficult to describe in accurate numerical value and it is more difficult to predict the nodes current status and priority based on the context information known as the historical beacons because of the dynamically changing topology, especially with an unexpected driver. For this reason, fuzzy logic is generally used here to divide the POS and $SPEED$ information into three levels respectively as shown in Fig.3(a) and Fig.3(b).The member function $\mu(SPEED)$ and $\mu(POS)$ describe the probability that the speed and position information belong to different levels. $ave(SPEED)$ and $ave(POS)$ are the vehicles average moving speed and the centre location.$min(SPEED)$ and $max(SPEED)$ are the minimum and maximum value of moving speed respectively, $min(POS)$ and $max(POS)$ are the positions of the last vehicle and the first vehicle in the communication range compared with the moving direction. In Fig.3(a), we take speed $S1$ as an example. It belongs to slow with a probability of 0.6 while medium with a probability of 0.4, and then the vehicle with speed $S1$ is thought of moving slow in the communication range. We can also imply that the speed S2 and S3 belong to medium while S4 belongs to fast. The same procedure may be easily adapted to divide the POS information into three levels in Fig.3(b). Table 1 show 9 possible status combined by different levels of *SPEED and POS* and the rules of priority prediction.



a.Membership function of position

b.Membership function of speed

Fig. 3. Membership function

4 bits are used to binarize the $SPEED$ and POS priorities of 9 statuses. Status 5 in Table 1 represents that a node with medium speed moving in the middle of the communication range should have the lowest priority $(1111)_2$ while status 1 or 9 represents the node that will leave the communication range soon should have the highest priority $(0000)_2$. From the 9 status of a vehicles $SPEED$ and POS information, we can deduce a status transition graph shown in Fig.4. Since status 2, 4, 6, 8 can transit to status 1 or 9 with only one step, their priority can be set as $(x_1x_2x_3x_4)_2$ in which x_i stands for either 0 or 1 randomly (i=1, 2, 3, 4). We take status 2 for instance, if all the x_i is 0, it means that we predict the vehicles with status 2 have reached status 1 in the current time, so its priority is set as $(0000)_2$. However, we can set $x_4 = \bar{x}_3$ when dealing with the priority of vehicle with status 3 or 7 in order to make it never get the highest priority $(0000)_2$ no matter $x_i(i = 1, 2, 3)$ is.

***ACC* Priority $w_i(ACC)$.** Acceleration or deceleration is an ordinary behavior of a vehicle node on the road but may be dangerous in an inappropriate scenario such as acceleration in a traffic jam or a sudden deceleration in highway.Those who accelerate or decelerate in an inappropriate scenario are potential threats to the road safety and should have a higher priority to access the channel to alert other nodes.

Table 1. Encoding rule of POS and ACC

$Status$	$\mu_{speed}(SPEED)$	$\mu_{pos}(POS)$	$BinaryPriority$
1	Back	Slow	0000
2	Middle	Slow	$x_1x_2x_3x_4$
3	Back	Slow	$x_1x_2x_3\bar{x}_3$
4	Back	Medium	$x_1x_2x_3x_4$
5	Middle	Medium	1111
6	Back	Medium	$x_1x_2x_3x_4$
7	Back	Fast	$x_1x_2x_3\bar{x}_3$
8	Middle	Fast	$x_1x_2x_3x_4$
9	Back	Fast	0000

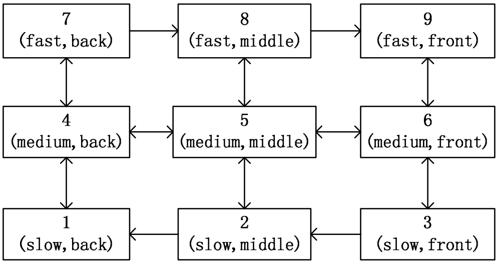


Fig. 4. Status transition graph

Since the ACC information which represents the behavior of each node varies a lot, we can also use fuzzy logic to divide it into 7 levels in Fig.5. From left to right, we name them level 0 to level 6. Now that we know whether a node is accelerating or decelerating, the next step is to detect whether the scenario is appropriate for this behavior.To a node, a simple way to define the scenario is to evaluate its relative speed and inter-distance compared with its immediate front neighbor. For example, a node with faster speed and smaller inter-distance compared with its immediate front neighbor lies in a scenario which is not suitable for acceleration while the scenario where a node with slower speed and larger inter-distance is not suitable for deceleration since its back neighbor may expect it to reduce the inter-distance. Fuzzy logic is used here as well to divide the relative speed and inter-distance of node i, for example, into three levels respectively as shown in Fig.6 both of which are derived in [12]. We assume the immediate front neighbor of node i is node j and t_s is the safety time for node i to reach node j with speed $SPEED_j$. The relative speed V_{ij} in Fig.6(a) can be calculated by $V_{ij} = SPEED_i - SPEED_j$. v_{max} is a predefined threshold value

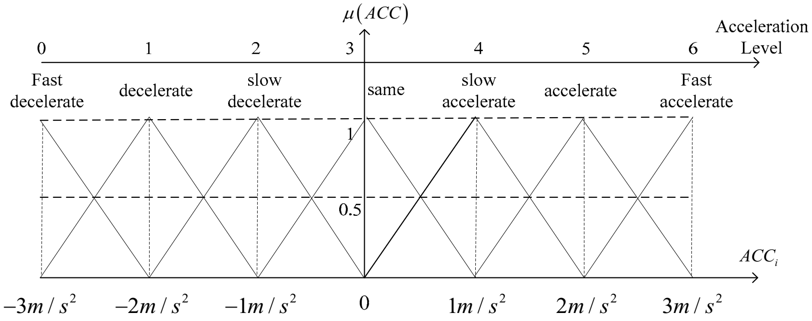
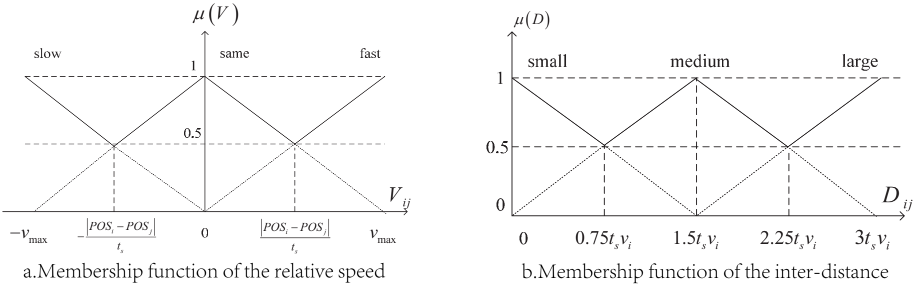


Fig. 5. Acceleration state



a.Membership function of the relative speed

b.Membership function of the inter-distance

Fig. 6. Membership function

which is the gap between the upper bound and the lower bound of the limited speed on the road. In Fig.6(b), the inter-distance $D_{ij} = POS_i - POS_j$ and v_i is short for $SPEED_i.\mu(V)$ and $\mu(D)$, which are the membership functions of the relative speed and inter-distance, define the probability that the relative speed and inter-distance belong to different levels.

The predicted $\mu(ACC)$ in Table 2 derived from [12] indicates the appropriate behavior of a node in the certain scenario which is defined by the combination of inter-distance $\mu(D)$ and relative speed $\mu(V)$. We take status 1 as an example, it means node i with small inter-distance between its immediate front neighbor, node j, and moves faster than j should have a fast deceleration (acceleration level 0) to avoid a potential crash.However, the unexpected driver may take a totally opposite move, such as a fast acceleration (acceleration level 6).When the actual behavior which is represented by the acceleration level l_a in Fig.5 differs greatly from the predicted one l_p in Table 2, it means that the node is so dangerous that it should have a higher priority to access the channel to alert other nodes. Since the larger difference between the actual acceleration level and the predicted one, the higher ACC priority is, we define a nodes ACC priority as $(x_1x_2x_3x_4x_5x_6)_2$, where $x_i = 0(i = 1, , |l_a - l_p|)$ and $x_j = 1(j = |l_a - l_p| + 1, \ldots, 6)$.6 bits are used

Table 2. Encoding rule of type

Status	$\mu_d(D)$	$\mu_v(V)$	*Predicted* $\mu_{acc}(ACC)$	*Predicted Level*
1	Small	Fast	Fast deceleration	0
2	Small	Fast	Deceleration	1
3	Small	Fast	Slow deceleration	2
4	Medium	Same	deceleration	1
5	Medium	Same	Same speed	3
6	Medium	Same	Acceleration	5
7	Large	Slow	Slow accelerate	4
8	Large	Slow	Accelerate	5
8	Large	Slow	Fast accelerate	6

here since the largest gap between the predicted level and the actual one is 6. The largest gap $|l_a - l_p|$ is 6 which makes node i get the highest ACC priority $(000000)_2$ while the smallest gap 0 which means the actual behavior correspond with the predicted behavior makes node i get the lowest priority $(111111)_2$.

Competing Priority $w_i(TS, TYPE, SPEED, POS, ACC)$. Since the nodes participated in the channel contention all lie in the one-hop communication range, they share the same context information, or the historical beacons, therefore, the sub-priorities predicted by each node is also the same (the random algorithms in calculating $w_i(POS, SPEED)$ in each node select the same pseudo-random generator and seed). By adjoining the four sub-priorities including $\lceil log_2 M \rceil - bit\ w_i(TS), 2 - bit\ w_i(TYPE), 4 - bit\ w_i(SPEED, POS)\ and\ 6 - bit\ w_i(ACC)$ in sequence, we can get an integrated $\lceil log_2 M \rceil + 12 - bit$ competing priority $w_i(TS, TYPE, POS, SPEED, ACC)$, where M is number of competing vehicle nodes.

3.2 Hamming Competition

The one with the highest competing priority discussed in part A will gain the opportunity to access the channel initiatively.Therefore we only need to find the one with highest priority, but not get the whole sequence list. Here we take the hamming competition as shown in Fig.7

The input $X = [x_1, x_2, x_j]^T = [1, 1, 1]^T$represents the lowest priority, the priority of node i $w_i(TS, TYPE, SPEED, POS, ACC)$which can be represented as $w_{ij}(i = 1, 2 \cdots M, j = 1, 2 \cdots 12)$ is the competing weights that link the inputs and the outputs in the forward sub-network. The output u_k calculated by (1)

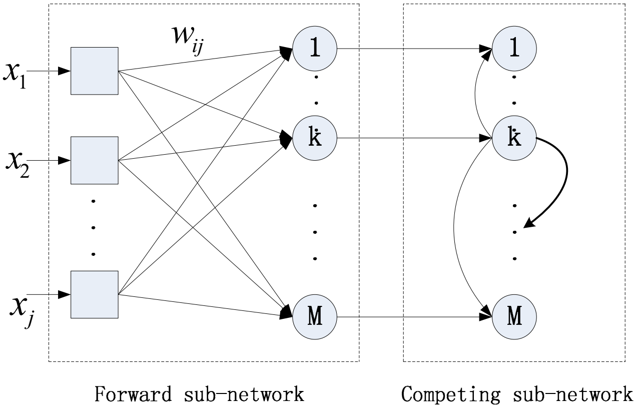


Fig. 7. A hamming network in a node

with the smallest value indicates that node k whose hamming distance is the largest is the competing winner, and then it gains the opportunity to access the channel.What need to be pointed out here is that if there're multiple nodes get the same smallest value, we choose the one with the smallest ID value as the competing winner.

$$u_i = \sum_{j=1}^{\lceil log_2 M \rceil + 12} w_{ij} x_j \tag{1}$$

Three advantages of adopting a hamming network model to select the highest priority vehicle node need to be demonstrated here. Firstly, the channel contention vehicle nodes and the competing priority can be directly and easily mapped to the competing nodes and the competing priority in a simple hamming competing network model. Secondly, the hamming network makes the calculation more structured and efficient especially with its memory function.Thirdly, the broadcast procedure after successfully accessing the channel can be mapped to the competing sub-network in a hamming network as shown in Fig.7 and this make this protocol more extensible.

3.3 Adaptive Back-Off

The competing winner node k in part B will access the channel initiatively and update the context information by broadcasting its beacon on the CCH immediately, which can be seen as the competing sub-network in Fig.7.All the other nodes will start a new round of channel contention with priority prediction based on the updated context information in part A, hence to realize the collision-free features in each contention.However, the complexity of VANETs make the winner in the hamming network still face some potential collision problems such as initiative contention and wait deadlock.We try to solve these two problems by adopting an adaptive back-off mechanism as shown in Fig.8 of which the part in the dashed frame is the detail implementations of adaptive back-off part in Fig.1.

Initiative Contention. Initiative contention means that both the nodes in the one-hop communication range and the ones in one-hop away may join the channel contention initiatively. It mainly caused by emergency.The following two situations may result in this problem. First, every node get a common winner k after the hamming competition except for node k_1 who thinks itself as the winner; Second, a node k_{M+1} from one-hop away rushes into the current communication area and try to access the channel. In both of the above two situations, k_1 and k_{M+1} will contend with winner node k initiatively and result in collisions without any back-off mechanism. We cannot use the global context information which include all the historical beacons broadcasted in the one-hop communication rangebecause node k_{M+1} doesnt know it at all and node k_1 will still get the same result since its context information hasnt been changed at all. Therefore, we try to avoid the potential collisions by analyzing the local context information which is only available to itself. A nodes local context includes the beacon message it has

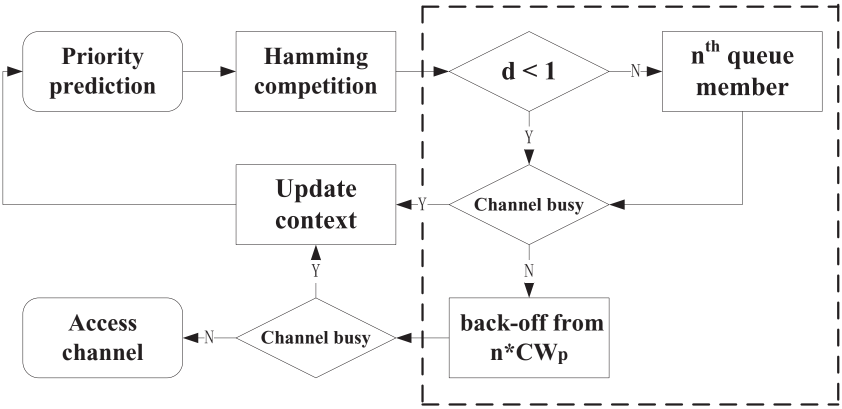


Fig. 8. Algorithm flow chart

just broadcasted, i.e., the historical status, and its current status. By comparing the change between these two statuses, we can get a nodes stableness d. The more the status changes, the less stable the node is, and the quicker the node should access the channel, therefore, the less contention window it should have.

The historical TS and $TYPE$ information will not change during this very time $\Delta t = t_{current} TS_i$, while the $SPEED$, POS and ACC may change more or less. A stable vehicular node will not change much since Δt is so short. However, the complexity of vehicular environment, especially the emergencies caused by the green hand driver or drunk driver, may make the vehicles status extremely unstable even in a short time Δt.

We take node i as an example, according to its current status including $SPEED_{current}, POS_{current}$ and $ACC_{current}$, we can get its positions in Fig.4 and Fig.5. Calculate the least transition steps d_1 and level gaps d_2 compared with that at time TS_i.We define stableness d as $\frac{1}{d_1+d_2}(d_1 + d_2 \neq 0)$, where d_1 is the transition steps and d_2 is the acceleration level gap between the historical status and current status. The less d is, the less stable the node is. In order to make unstable nodes to access the channel as soon as possible, we assign the node n CW_p as their contention window, where $n = d$,CW_p is the initial contention window in IEEE 802.11p.

Wait Deadlock. The complexity of vehicular environment leads to the probability of losing beacons, therefore, the global context information based on which to predict priority by each node may differ from each other. This may cause the following scenario: each contention node gets a common winner k though their hamming networks respectively except that k itself thinks k is the winner. When all the nodeswait for k to access the channel, however node k is waiting for k to access the channel, resulting in the wait deadlock which also causes a great waste of channel resource.

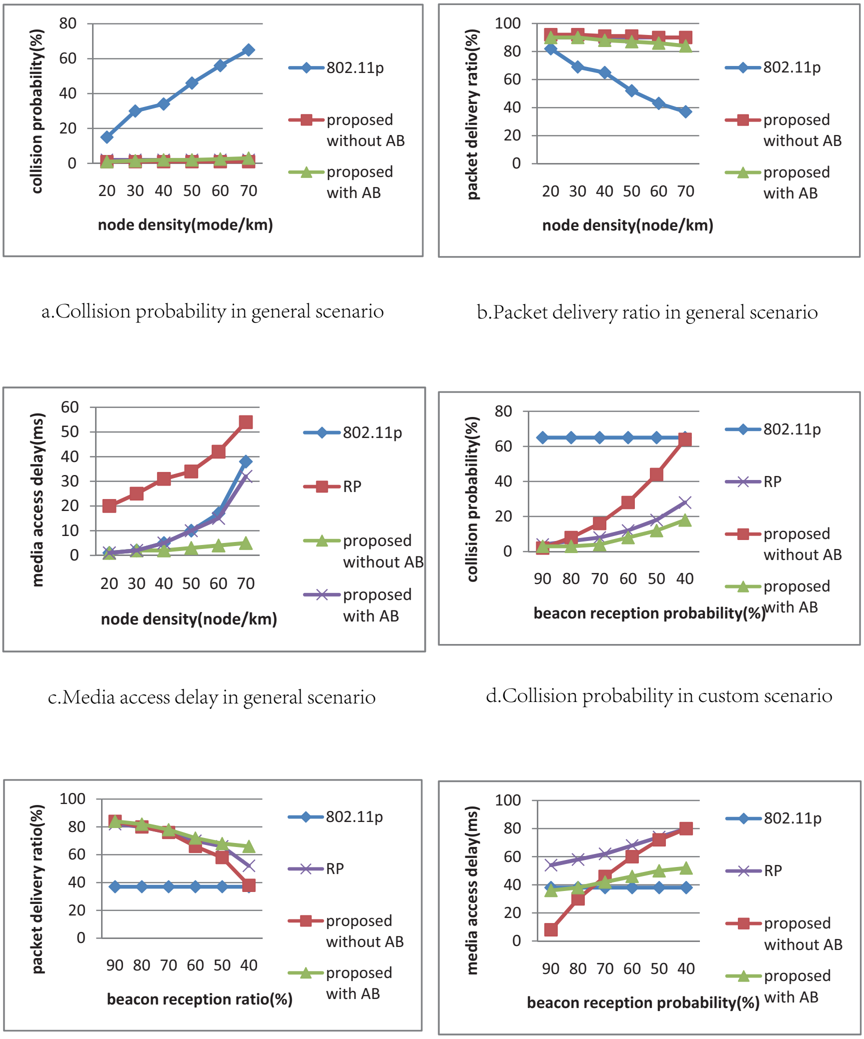
In order to solve the deadlock problem, we introduce a competing queue arranged descendingly according to the hamming distance in part B or ascendingly according to u_i in (1). The node in the 1^{st} place of the queue is the one with the smallest u_i and it randomly selects its back-off time CW_p. The node in the 2^{nd} place of the queue sets its contention window as $2CW_p$, by parity of reasoning the node in the i^{th} place of the queue sets its contention window as iCW_p where $i \leq \lceil log_2 M \rceil + 12$ since only $\lceil log_2 M \rceil + 12$ priorities can be differentiated in part B. The worst situation is that node j waits for node $(j + 1) mod M$ to access the channel and each node thinks itself lies in the end or the $\lceil log_2 M \rceil + 12^{th}$ place of the queue($1 \leq j \leq M$, M is the number of contention nodes). Under this circumstance, the proposed protocol will degrade as the traditional 802.11p protocol but with a contention window of $\lceil log_2 M \rceil + 12$CWp.

With the adaptive back-off mechanism, the node whose back-off time expires first will access the channel and start transmission. The broadcasted beacon will update the global context information and another round of channel contention will begin with this context-aware method.

4 Evaluation

In this section, three metrics including collision probability, packet delivery ratio, and media access delay are used to evaluate the performance of 802.11p, Reservation Protocol[7], our proposed MAC protocol without Adaptive Back-off mechanism (proposed without AB) and the one with Adaptive Back-off (proposed with AB) under various node density and context acknowledgementscenarios. Collision probability is calculated as the number of collisions (which cause packet drops) detected by the receiving antenna divided by the total number of messages sent in the communication range. The packet delivery ratio is the number of messages received to the total number of messages sent within the communication range. The medium access delay is defined as the time gap between the time a packet arrives at the MAC layer and the time the packet actually starts transmission. All three metrics are accumulated and averaged all nodes during the entire simulation.

For these simulations, we have turned to NS-3[13] and a straight highway model[14]was used to define a 3-kilometer unidirectional highway with two lanes. Automatically-generated vehicles are injected to the highway with speed ranging from 20 - 30m/s and inter-distance of two neighboring ones ranging from 15 - 50 meters. We assume that all vehicles are synchronized to the control channel interval through a GPS system to send their beacon messages and in the initial scenario, status messages have already been exchanged with each other in their transmission range which takes any of the value 250, 300, 350 and 400 meters. In order to satisfy the demand of this highway scenario when the vehicles running on it are saturated, we set M as constant 64, which means 6 bits are used to represent TS priority. The initial simulation parameters are set as follow: Modulation (QPSK), Transmission frequency (5.9GHz), Bit rate(6Mbps), Packet size (6Kb), Data rate (25Hz), Safety timets(3s), CWp(31), Slot time($16 \mu s$), SIFS($32 \mu s$).



a.Collision probability in general scenario

b.Packet delivery ratio in general scenario

c.Media access delay in general scenario

d.Collision probability in custom scenario

e.Packet delivery ratio in custom scenario

f.Media access delay in custom scenario

Fig. 9. Experimental results

The parameters of Reservation Protocol is set as the same as that in [7]. Fig. 9(a) shows our proposed MAC protocols achieves the same goal of near collision-free as that of RP and has a considerably lower collision probability than IEEE 802.11p protocol. However, compared with RP the collision probability of the proposed protocol with AB is still a little sensitive to node density since all the nodes mobility characteristics tend to be the same with the increase of the node density, more than one winner will be obtained through the hamming networks.

During the adaptive back-off phase, the same mobility results in a same high stableness which also make the collisions increase.

Fig. 9(b) presents the packet delivery ratio, which is often used to characterize communication reliability. The performance of 802.11p deteriorates as the density increases while RP and the proposed protocols still preserve a relatively high delivery ratio. It is reasonable that the delivery ratio of the proposed protocol with AB is also a little sensitive to the density since the collision probability increases a little with the density.

Fig. 9(c) shows RP has a considerably higher media access delay, 802.11p and our proposed protocol with AB preserve relatively low whilethe proposed protocol without AB is almost zero delay. What contributes most to this featureis the Adaptive Back-off mechanism. We have demonstrated it in Fig.2 that the priority prediction and hamming competition can be processed during the time the node transmits on the SCH after finishes beaconing, which will not introduce extra time overhead.Therefore, the node adopts the protocol without AB will access the channel immediately withoutmuch delay while the one with AB wont access the channel until the back-off time expires.

The simulation scenario discussed above is relatively ideal with the assumption that each vehicle in the communication range acknowledges the global context information and the entirevehicles move regularly on the highway without any disturbance, the simulation results shown above indicate that our proposed protocol without AB gainsmuch better performance instead.However, the real vehicular environment is much more complicated, especially the beacon loss caused by nonideal channel or signal attenuation and the unexpected emergencies caused by unpredicted drivers behavior will both affect the performance of the above four protocols.In order to verify the necessity and validity of the Adaptive Back-off mechanism in Part B, we set up a custom scenario to reflect the complexity discussed above. To simulate the beacon loss, we try to make the global context information of each node different from each other by setting abeacon reception probability. The node density is set as 70node/km and 10% percent of all the nodes running on the highway are chosen to be assigned with a random speed and a corresponding position every 10ms so as to simulate the unpredicted drivers behavior. Then we evaluate the performance of the four protocols again when each node receives beacons from its neighbors with a probability of 90%, 80%, 70%, 60%, 50% and 40% respectively.

Fig. 9(d) reflects the collision probability varies as the beacon reception probability.The working procedure of 802.11p doesn't rely on beacon information, therefore itsperformanceincluding collision probability, packet delivery ratio and access delay won't be affected by beacon loss.The less the beacon reception probability, the less global context information can be used by a node to predict priorities, and the higher collision probability will be caused by the proposed protocol without AB which entirely relies on the completeness of the global context information. RP is also sensitive to reception probability since its prediction algorithm also depends on the beacon information. Our proposed protocol with AB gains the lowest collision probability by adaptively adjusting CW of the potential contenders.

Fig. 9(e) shows the packet delivery ratio of the four protocols with different beacon reception probability. The reason of the drop of the packet delivery ratio is the same as the reason of the collision probability. Our proposed protocol with AB always maintains a relatively high packet delivery ratio.

Fig. 9(f) reflects the increase trend of the protocols average media access delay. Since its complicated reservation mechanism, RP still remains a relatively high access delay. What surprises us most is the average delay of our proposed protocl without AB surges from almost zero to a relatively high level. The main reason is the increase of the collision probability. Our proposed protocol with AB also gains a little higher delay than 802.11p since the CW parameter n changes too often.

5 Conclusion and Future Work

In this paper, we have proposed a context-aware MAC protocol for VANETS. By forming a simple hamming network of which the neuron cells are all the nodes in the communication range and the competing weights are their priorities predicted based on the global context in each node, all the nodes will get a common competing winner though their hamming networks. However, problems such as beacon loss or unpredicted emergencies caused by the complexity of the vehicular environment make potential collisions still exist. Experimental results demonstrate the proposed MAC protocol reduce the collision probability considerably and improve the transmission reliability while keeping a low media access delay. In the future, we intend to explore more details of the context information derived from historical beacons and the fuzzy logic used to prioritize each node can also be improved to make the predicted priorities more accurate and differentiated.

Acknowledgment. This work is supported by the the NSF of China under grant 61272485.

References

1. Hiertz, G., Denteneer, D., Stibor, L., Zang, Y., Costa, X.P., Walke, B.: The IEEE 802.11 universe. IEEE Communications Magazine 48(1), 62–70 (2010)
2. Strom, E.G.: On Medium Access and Physical Layer Standards for Cooperative Intelligent Transport Systems in Europe. Proceedings of the IEEE 99(7), 1183–1188 (2011)
3. Jian, Y., Chen, S.: Can CSMA/CA networks be made fair? In: Proceedings of the 14th ACM International Conference on Mobile Computing and Networking, pp. 235–246 (2008)
4. Acatauassu, D., Couto, I., Alves, P., Dias, K.: Performance Evaluation of Inter-Vehicle Communications Based on the Proposed IEEE 802.11 p Physical and MAC Layers Specifications. In: The Tenth International Conference on Networks, pp. 170–174 (2011)

5. Balon, N., Guo, J.: Increasing broadcast reliability in vehicular ad hoc networks. In: Proceedings of the 3rd International Workshop on Vehicular Ad hoc Networks, pp. 104–105 (2006)
6. Lans, H.: Position indicating system. Google Patents (1996)
7. Zhang, S., Cahill, V.: Towards collision-free medium access control in vehicular ad-hoc networks. In: Proceedings of the Eighth ACM International Workshop on Vehicular Inter-Networking, pp. 83–84 (2011)
8. Yu, F., Biswas, S.: A self reorganizing MAC protocol for inter-vehicle data transfer applications in vehicular ad hoc networks. In: 10th International Conference on Information Technology (ICIT 2007), pp. 110–115 (2007)
9. Lu, N., Wang, X., Wang, P., Lai, P., Liu, F.: A distributed reliable multi-channel MAC protocol for vehicular ad hoc networks. In: IEEE Intelligent Vehicles Symposium, pp. 1078–1082 (2009)
10. Wang, Y., Ahmed, A., Krishnamachari, B., Psounis, K.: IEEE 802.11 p performance evaluation and protocol enhancement. In: IEEE International Conference on Vehicular Electronics and Safety, ICVES 2008, pp. 317–322 (2008)
11. Karamad, E., Ashtiani, F.: A modified 802.11-based MAC scheme to assure fair access for vehicle-to-roadside communications. Computer Communications 31(12), 2898–2906 (2008)
12. Abdel Hafeez, K., Zhao, L., Liao, Z., Ngok-Wah Ma, B.: Clustering and OFDMA-based MAC protocol (COMAC) for vehicular ad hoc networks. EURASIP Journal on Wireless Communications and Networking 2011(1), 1–16 (2011)
13. Henderson, T.R., Lacage, M., Riley, G.F., Dowell, C., Kopena, J.B.: Network simulations with the ns-3 simulator. In: SIGCOMM Demonstration (2008)
14. Arbabi, H., Weigle, M.C.: Highway mobility and vehicular ad-hoc networks in ns-3. In: Proceedings of the 2010 Winter Simulation Conference (WSC), pp. 2991–3003 (2010)

A Balance Storage Nodes Assignment for Wireless Sensor Networks

Zhigang Li[1], Geming Xia[2], Weiwei Chen[1], and Qing Li[1]

[1] College of Command Information System, PLA Univ. of Sci. and Tech., China
[2] School of Computer, National University of Defense Technology, China

Abstract. A balance storage nodes assignment problem is proposed for wireless sensor networks in order to save history data as much as possible when the network cannot connect to the sink node(s). Its aim is to assign storage nodes as evenly as possible to each source node, while it can minimize the storage path cost of the whole network. This problem can be reduced into a classic assignment problem by adding "dummy" source nodes, which is not easy to solve through the existing centric algorithms in sensor networks. Then two algorithms (Random Greedy Algorithm and Voronoi Graph Based Algorithm) are proposed in order to solve this problem. These two algorithms can be implemented distributed for the wireless sensor networks. The simulation implements and compares the performance of these two algorithms. The result shows that the Voronoi Graph Based Algorithm can satisfy the load balance storage for all source nodes and achieve approximate minimal storage path cost of the whole network.

Keywords: sensor networks, balance storage, assignment problem.

1 Introduction

In many applications of wireless sensor networks, network connection failure and packets loss often happen because of individual nodes failure and routing congestions [1]. In such cases, the sink node(s) cannot be guaranteed to receive the collecting data from the sensor nodes. So it is necessary to store the history data in sensor nodes for a while, in order to send them to the sink when the network gets resilience [2], however, the storage capacity of individual sensor node is limited and cannot support larger scale local storage. In sensor networks, not all the nodes are responsible for sensing and collecting data. A node is called source node if it can sense event data, otherwise it is called storage node if it is a redundant node and is mainly for storing data. The source node can disseminate the history data to some redundant storage nodes [3,4], in order to store more data in-network.

For the in-network storage of wireless sensor network, although the total data account of the whole network is a very important factor, the data balance of each source node is another key point. Because, in sensor networks, the data of different sources have much correlation, therefore it is necessary to cooperate with all the data of all the distributed source nodes in order to mine the interested events [5]. In other words, it is

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 151–162, 2013.

incomplete to describe an event only by part of source nodes' data in some applications. Although source nodes can store their data to other storage nodes, however, if source node S_1 is assigned less storage nodes than source node S_2, it is possible that S_1's data get lost more than S_2. On the other hand, for source node S_2, even it has more data stored than S_1, however, the node S_2's data cannot describe the whole state of the event without S_1's data participation. Then the data of node S_2 become useless and the storage nodes for storing data of node S_2 waste energy. It means the storage balance for all source nodes is very important. The first aim of this work is to assign storage nodes to each source node equally as possible as much.

As the source node needs multi-hop transmission for sending data to the storage nodes, the overall hop distances between source nodes and storage nodes should be as short as possible, in order to minimize transmission energy cost.

The existing in-network storage of wireless sensor network includes two categories, coding-based storage and data-copy-based storage. The source node can divide its data into several segments, and these segments can be coded with other source nodes' data segments in coding-based storage approach. In coding-based storage approach, if one storage node fails, its data can be recovered by decoding from other storage nodes. In literature [6,7], fountain code is used for disseminating k source nodes' data to other storage nodes randomly, by XOR operation in every storage node. The original k data can be decoded by visit and obtain arbitrary $(1+\varepsilon)k$ storage nodes with high probability. The classic data-copy-based storage approach is GHT [4]. The idea of GHT is as follows. First the data's type or Key is mapping to a location in the network by a Hash function. Then the node near this location is assigned as a storage node for this source node. The aim of GHT is for data discovery, without considering load balance and location-free condition. Another related work is Voronoi graph based nodes clustering approach. In this approach, the nodes number in each Voronoi cell is different, so it cannot solve the load balance assignment problem of this work. Our work refers and improves the Vornoi graph based approach, in order to solve and implement load balance storage assign problem.

In our solution, the in-network storage is a complementation for a sink-based wireless sensor network, while the sink is failed to retrieve data. When the sink is resilience, all the source nodes and storage nodes should send their data to the sink automatically.

2 Problem Statements

This work assumes that one storage node can only service for storing data from one source node, in order to avoid collision and interference of multiple source nodes' data in one storage node, which increases much management overhead. Then it is supposed that there are k source nodes and n storage nodes in a sensor network. The aim of this work is to divide the storage nodes to k part evenly, and assign each part to one of the k source nodes, while the overall path length from source node to storage node is minimal. This problem can be described formally as follows.

Definition 1: Storage Nodes Assignment Problem
The sensor network can be considered as a undirected graph G=(V,E), where V=$C\cup S$, C={v_1,v_2,…,v_k}, S={ v_{k+1},v_{k+2},…,v_{k+n} }, k<n. The edge set $E\subseteq V\times V$. The storage nodes assignment problem is to partition the graph G into k subset $P_{1,}P_2,\cdots P_k$,s.t.,

(1) $\bigcup_{i=1}^{k} P_i = V$, and $P_i \cap P_j = \Phi$, $i\neq j$;

(2) $v_i \in P_i$, $1\leq i\leq k$;

(3) $\left|(\mid P_i\mid - \mid P_j\mid)\right| \leq 1$, $\mid P_i\mid$ is the number of nodes subset P_i , $i\neq j$, $1\leq i,j\leq k$;

(4) $\min\left(\sum_{1\leq i\leq k}\delta(P_i)\right)$, where $\delta(P_i)$ is the sum of the path length from one source node to one storage node, i.e., $\delta(P_i) = \sum_{1\leq r\leq\mid P_i\mid} h(v_i, v_r^{'})$, $v_r^{'}\in P_i$, $h(v_i,v_r^{'})$ is the hop distance between $v_i, v_r^{'}$. For expression simplicity, $\sum_{1\leq i\leq k}\delta(P_i)$ is also called storage path cost, and the storage nodes assigned to v_i in P_i is called slave storage nodes or slave nodes of source node v_i.

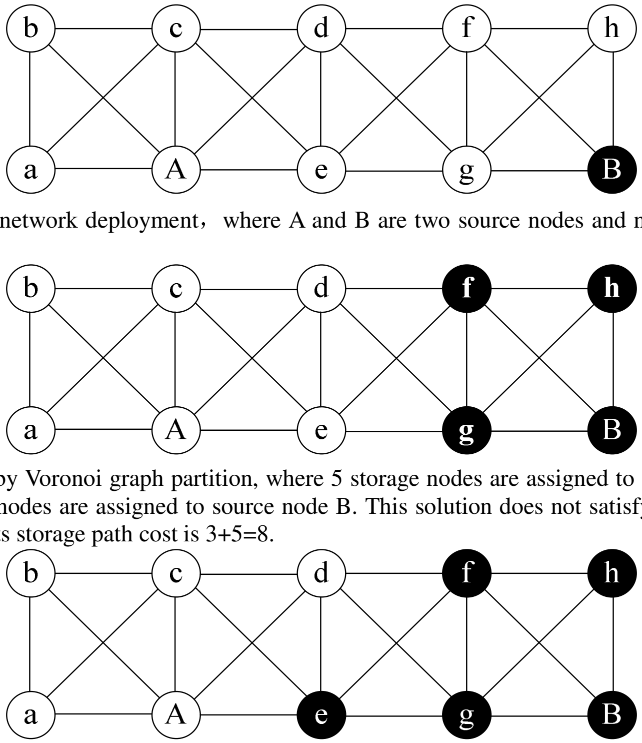
An Simple Example
Figure 1 gives an example to illustrate the storage nodes assignment problem. The figure 1.a shows a network with 10 nodes, where A and B are two source nodes and others are storage nodes. The figure 1.b shows an assignment solution by Voronoi graph partition method [6], where the 1-hop neighbors are assigned to A and B respectively. In this assignment solution, A has 5 slave nodes and B has 3 slave nodes. The storage path cost is 5+3=8, however, this assignment approach does not satisfy the storage balance requirement, the third item in the above definition. In figure 1.c, storage node e is assigned to B, which makes A and B both have 4 slave nodes and the storage path cost is 4+5=9. Because the load balance is the first primitive objective, the storage path cost of the second approach is optimal at the storage balance constraint.

3 Problem Analysis

This section analyses the above storage nodes assignment problem and induces it to a classic assignment problem.

3.1 Bipartite Graph Balance Partition Problem

The storage nodes assignment problem can be transformed to a weighted bipartite graph partition problem as follows,



(a) A 2×5grid network deployment，where A and B are two source nodes and node a to g are storage nodes.

(b)A solution by Voronoi graph partition, where 5 storage nodes are assigned to source node A and 3 storage nodes are assigned to source node B. This solution does not satisfy load balance requirement. Its storage path cost is 3+5=8.

(c)If node e is assigned to source node B, it is a load balance partition solution. The storage path cost is 5+4=9 for this solution.

Fig. 1. Example of storage nodes assignment

Definition 2. Bipartite Graph Balance Partition Problem

$G=(V_1 \bigcup V_2, E)$ is a complete bipartite graph, $V_1 \bigcup V_2$ are nodes set, where $V_1=\{v_1,v_2,...,v_k\}, V_2=\{v_{k+1},v_{k+2},...,v_{k+n}\}$, *k*<*n*. *E* is edge set, each edge e_i has a weight w_i.

The set V_2 are divided to *k* part $P_{1,}P_2,\cdots P_k$, s.t.,

(1) $\bigcup_{i=1}^{k} P_i = V_2$, and $P_i \bigcap P_j = \Phi$, $i \neq j$;

(2) $\max\limits_{1\le i\le k, 1\le j\le k} \left(\| P_i | - | P_j \| \right) \le 1$, $| P_i |$ is the nodes number of subset P_i;

(3) $\min\left(\sum\limits_{1\le i\le k} W(P_i) \right)$, where $W(P_i)$ is the weight sum from v_i to all the nodes of P_i, i.e., , $v_r^{'} \in P_i$,and $w(v_i, v_r^{'})$ is the weight of edge $v_i, v_r^{'}$.

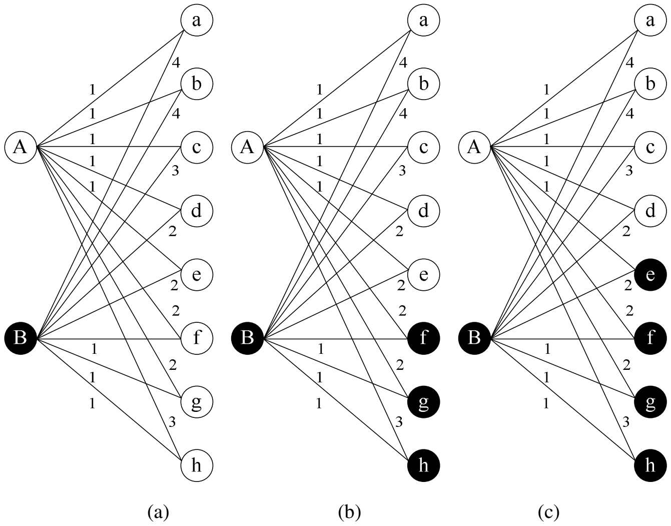


Fig. 2. Bipartite graph based nodes assignment

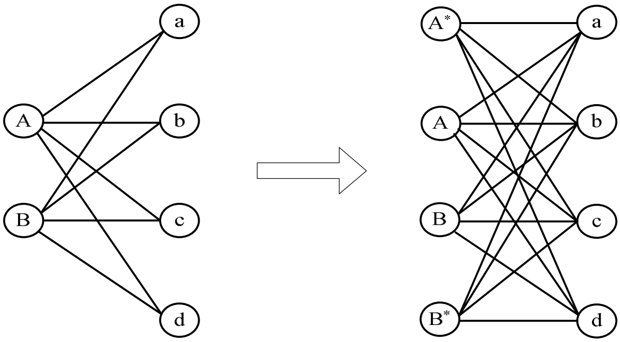


Fig. 3. Bipartite graph balance partition problem to complete bipartite graph weighted balance matching problem

Figure 2 is the example transformed from figure 1, where V_1={A,B}, V_2={a,b,c,d,e,f,g,h}. If V_2 is partitioned to P_1={a,b,c,d}, P_2={e,f,g,h}, it is a optimal solution.

3.2 Balance Minimal Weight Matching Problem

The balance bipartite graph partition problem can be transformed to complete bipartite graph balance minimal weight matching problem by adding “dummy” source nodes. Two cases are considered as follows.

First we suppose k is an exact divisor of n, and then each source node v_i in V_1 is copied n/k-1 dummy nodes. The weight of each dummy node to the node of V_2 equals the weight of v_i to the node of V_2. As figure 3.a shows a complete bipartite graph balance partition problem of $K_{2,4}$. Source node A and B both need two storage nodes. In figure 3.b, A* and B* are dummy source nodes of A and B respectively, by copying from A and B. The weights from A* and B* to a, b, c, d equals the weights from A and B to a, b, c, d respectively. A, B, A*, B* all need one storage node.

Second, if k is not an exact divisor of n, in order to satisfy the constraint $\max_{1\le i\le k, 1\le j\le k}(||P_i|-|P_j||)\le 1$, so each source node has $\lceil n/k \rceil$ storage nodes at most and $\lfloor n/k \rfloor$ storage nodes at least. We copy k_1 source nodes $\lfloor n/k \rfloor$-1 dummy nodes and other k_2 source nodes $\lceil n/k \rceil$-1 dummy nodes, where k_2=n mod k.

$v=\{v(1),...,v(k)\}$, $v(k)\in\{0,1\}$, and $\sum_{1}^{k} v(k)=k_2$, $v(i)$ represents different cases of dummy nodes of source node i. If $v(i)=1$, it shows node i has $\lceil n/k \rceil$-1 dummy nodes; if $v(i)=0$, it shows node i has $\lfloor n/k \rfloor$-1 dummy nodes.

By (1) and (2), it shows the bipartite graph balance partition problem is equivalent to assignment problem.

The Hungarian Method is a classic solution for the assignment problems. But generally the Hungarian Method is a centric algorithm and its time complexity is $O(n^3)$, which is not appropriate for wireless sensor network.

4 Heuristic Algorithm Design

First we give two heuristic algorithms according to the sensor networks' specialty and then discuss how to implement one of them in the distributed solution.

4.1 Random Greedy Algorithm

The idea of random greedy algorithm is that the source nodes try to select the nearest storage nodes. First node v_i randomly is selected from V_1. The unassigned nodes in V_2 are ascending sorted by the weights to v_i. Then the front $\lceil |V_2|/|V_1| \rceil$ nodes subset P$_i$ of V_2 are assigned as storage nodes of v_i. At the beginning of each loop, V_1= V_1-v_i, V_2= V_2-P_i.

The main part of this algorithm is the sorting process for the nodes of V_2. The time complexity is $O(n\log n)$ .The greedy algorithm may meet local maximal problem that some source nodes have selected nearer nodes as their slave nodes, which makes other source nodes have to select far storage nodes as their slave nodes. Then the overall storage path cost increases.

Algorithm 1. Greedy Algorithm

```
Input: nodes set $V_1$ ,$V_2$ ,$k$=|$V_1$|.
Output: $P_{1,}P_2,\cdots P_k$ .
1. If $V_1$ is empty, return; else go to 2;
2. Randomly select $v_i \in V_1$;
3. For all the nodes in $V_2$, sorting them according to $e(v_i,u)$ increasingly, then a temporary set TempSet={$u_1,u_2,\ldots,u_n$} got, $n$=|$V_2$|;
4. P$_i$={ $u_1,\ldots,u_l$}; $l=\lceil |V_2|/|V_1| \rceil$;
5. $V_1= V_1-v_i$, $V_2= V_2-P_i$;
6. Goto 1;
7. Output $P_{1,}P_2,\cdots P_k$ .
```

4.2 Voronoi Graph Based Algorithm

The Voronoi Graph partition method can be used to assign storage nodes to sources nodes. But the Voronoi cells include different number of storage nodes as the source nodes generate dispersedly, randomly and irregularly. Intuitively, the source nodes with more storage nodes by the Voronoi Graph partition method need hand out some storage nodes to the source nodes with less storage nodes, in order to achieve balance. From this idea, a new heuristic algorithm is designed as follows.

Heuristic Rule: If the number of sensor nodes Voronoi cell of source node S_i is larger than $\lceil |V_2|/|V_1| \rceil$, the source node S_i needs to select such sensor nodes which are far to other source nodes, in order to prevent other source nodes from selecting far sensor nodes as their storage nodes. The detail selecting rule is as follows. First, the storage path cost of each sensor node in S_i Voronoi cell to S_i is computed. Then all the sensor nodes in S_i Voronoi cell are sorted according to their storage path cost gains decreasingly. At last the former $\lceil |V_2|/|V_1| \rceil$ sensor nodes are assigned as storage nodes of source S_i. The storage path cost gain is defined as follows,

$$gain(v) = \min\{H_j(v) \mid j \in [1,k], j \neq i\} - H_i(v),$$

where $H_i(v)$ and $H_j(v)$ are hop distance from node v to source S_i and S_j. The storage path cost gain is the plus storage path cost when source S_i does not select v as slave node and the second nearer source node selects v as slave node, the. So the source node S_i should select storage node with bigger storage path cost gain, in order to minimize the overall storage path cost of the network. Algorithm 2 gives the detail of Voronoi Graph based algorithm.

The time complexity of this algorithm is also $O(n\log n)$. But this algorithm is easy to implement distributed.

Algorithm 2. Voronoi graph based algorithm

```
Input: nodes sets $V_1$, $V_2$, $k$=|$V_1$|.
Output: $P_{1,}P_2,\cdots P_k$.
1. Each Voronoi Graph cell of each node in $V_1$ is computed according the path hop
   distance between nodes in $V_1$ and $V_2$.
2. Sorting nodes in $V_1$ according to nodes number in each Voronoi Graph cell de-
   creasingly.
3. Find $s_i$ with largest nodes number of Voronoi Graph cell, and mark the storage
   nodes in Voronoi Graph cell of $s_i$; sorting the nodes in Voronoi Graph cell of $s_i$
   according to their storage path cost gain decreasingly
   $gain(v) = \min\{H_j(v) \mid j \in [1,k], j \neq i\} - H_i(v)$.
   The sorted temporary set is TempSet={ $u_1,u_2,\ldots,u_n$ }, $n=|Vo(s_i)|$;
4. $P_i$={ $u_1,\ldots,u_l$}; $l=\lceil |V_2|/|V_1| \rceil$;
5. $V_1= V_1-v_i$, $V_2= V_2-P_i$
6. Goto 1.
7. Output $P_{1,}P_2,\cdots P_k$.
```

4.3 Distributed Voronoi Graph Based Algorithm

The above two algorithms are both centralized algorithm, which cannot be used to the sensor network directly. We discuss how to implement a distributed Voronoi Graph based Algorithm in this section. In distributed algorithm, the key step is about how the source node to obtain the number of sensor nodes in its Voronoi cell and to which source node a sensor node should be assigned. The number of source nodes, the number of storage nodes and the number of slave nodes of each source node could be computed at the initialization of the network [3].

The basic idea of distributed Voronoi Graph based algorithm is that each storage node compute its path storage cost gain and send its gain value to each source nodes in distributed way. Each source node sorts the received gain values decreasingly. The $\lceil |V_2|/|V_1| \rceil$+1 gain value g is selected as a filter to be sent to other storage nodes by source node S_i. When a storage node receives the value of g, it compares its gain value with g. If its gain value is bigger than g, this storage node is assigned to source node S_i as a slave node.

5 Performance Evaluation

We use MATLAB to evaluate the performance of the proposed algorithms. First, we set up a simple grid network with 49 nodes to test the feasibility and differences of two algorithms. Then on an irregular network with 600 nodes, the storage path cost performance is tested.

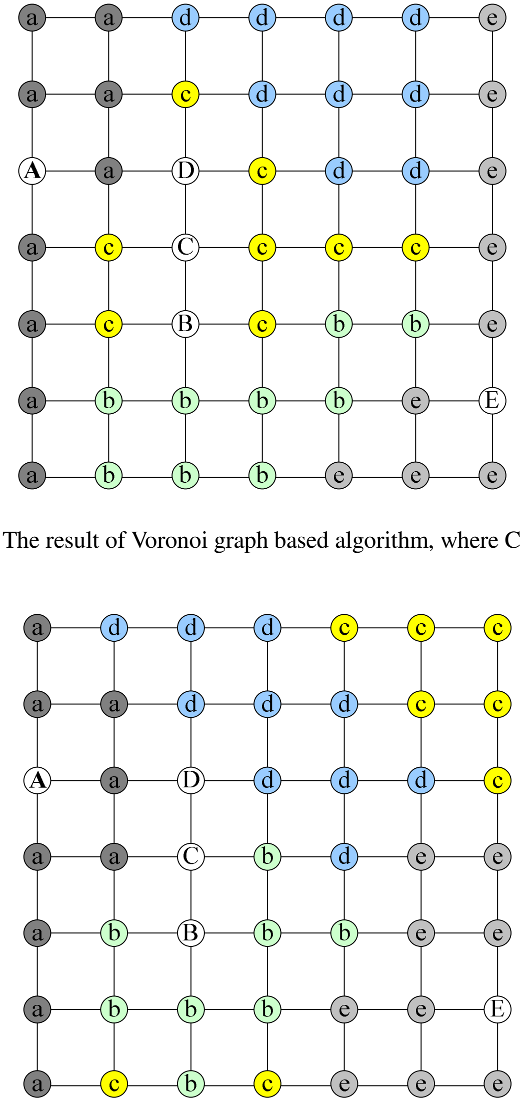
5.1 Storage Balance Simulation

In a 7×7 grid network, 5 source nodes A, B, C, D, and E are generated randomly. Other 44 nodes are as storage nodes. The simulation shows both the greedy algorithm and Voronoi Graph Based algorithm can achieve network nodes balance partition and storage balance. Figure 4.a shows the running result of Vornoid graph based algorithm, where A, B, D, and E get 9 storage nodes as slave nodes and C gets 8 storage nodes as slave nodes. Figure 4.a shows that the slave nodes of each source are distributed closely and nearly. Figure 4.b shows the running result of greedy algorithm, likely, A, B, D, and E get 9 storage nodes as slave nodes and C gets 8 storage nodes as slave nodes. For the running result of greedy algorithm, however, the slave nodes of source node C are distributed at different areas of network and far from source node C. This is because the source node C is assigned storage nodes left by source node A, B, D, E selecting their own slave nodes greedily and optimally according to the greedy algorithm. Source node C actually has no other chance to select other storage nodes.

Figure 5 shows the storage path cost distribution of each source nodes of two algorithms respectively from another point of view. Figure 5.a shows the Voronoi Graph based algorithm does not only guarantee balance assignment of storage nodes, but also achieve balance storage path costs of all the source nodes. Figure 5.b shows the storage path cost of source C is much higher than other source nodes by running greedy algorithm.

5.2 Storage Path Cost Test

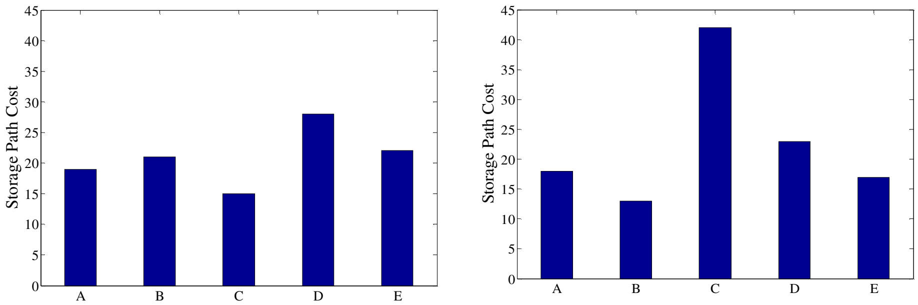
600 sensor nodes are randomly deployed in an irregular sensor network. At the first step, 5 source nodes are generated randomly. Then new source node will be generated one by one until 14 source nodes. Figure 6 shows, as the number of source nodes increasing, the total storage path cost decreases both for Voronoi graph based algorithm and greedy algorithm. This is because as the number of source nodes increase, storage nodes should be divided into smaller group, and can be assigned to nearer source node. When the source nodes number is 8, the figure 6 shows that its total storage path cost is larger than the case with 7 source nodes. This phenomenon also indicates that the greedy algorithm’s limitation, which may make some source node selecting far storage nodes, even the source node number increases. The Voronoi Graph based algorithm is better than the greedy algorithm on storage path cost, which is more close to the lower bound.



(a) The result of Voronoi graph based algorithm, where Cost=105

(b) The result of greedy algorithm, where Cost=113

Fig. 4. Result of load balance test



(a) Storage path cost distribution of Voronoi graph based algorithm

(b) Storage path cost distribution of greedy algorithm

Fig. 5. Storage path cost distribution from source nodes to storage nodes

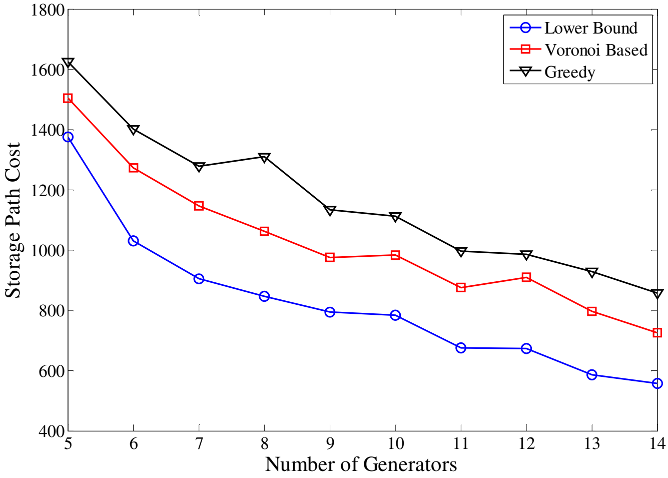


Fig. 6. Storage path cost test

6 Conclusion

A balance storage nodes assignment problem is proposed for wireless sensor networks, and this problem can be reduced into a classic assignment problem. Two algorithms are proposed for solving this problem, greedy algorithm and Voronoi graph based algorithm. The performance of two algorithms is test by simulation.

Acknowledgment. The research was partially supported by the National Natural Science Foundation of China under Grant No. 61070201 and No. 61070199, the Natural Science Foundation of Jiangsu Province under Grant No. BK2011119. We would also like to thank the anonymous reviewers for their constructive comments.

References

1. Taddia, C., Mazzini, G.: On the retransmission methods in wireless sensor networks. In: Proceedings of IEEE 60th Vehicular Technology Conference, Mazzini, Italy, pp. 4573–4577 (2004)
2. Petit, L., Nafaa, A., Jurdak, R.: Historical data storage for large scale sensor networks. In: Proceedings of the 5th French-Speaking Conference on Mobility and Ubiquity Computing, NY, USA. ACM (2009)
3. Aral, B., Gunopulos, D.: Reliable hierarchical data storage in sensor networks. In: Proceedings of 19th International Conference on Scientific and Statistical Database Management, Banff, Alberta, Canada (2007)
4. Sylvia Ratnasamy, B.K., Yin, L., Yu, F., Estrin, D., Govindan, R., Shenker, S.: GHT: A Geographic Hash Table for Data Centric Storage. In: Proceedings of the 1st ACM WSNA, Atlanda, GA, USA, pp. 78–87 (2002)
5. Ari, I., Çelebi, Ö.F.: Finding event correlations in federated wireless sensor networks. In: Procedings of 7th International Wireless Communications and Mobile Computing Conference, IWCMC (2011)
6. Aly, S.A., Kong, Z., Soljanin, E.: Fountain Codes Based Distributed Storage Algorithms for Large-scale Wireless Sensor Networks. In: Proceedings of International Conference on Information Processing in Sensor Networks, St. Louis, Missouri, USA, pp. 171–182 (2008)
7. Lin, Y., Liang, B., Li, B.: Data persistence in large-scale sensor networks with decentralized fountain codes. In: Proceedings of 26th IEEE INFOCOM, Alaska, USA, pp. 1658–1666 (2007)
8. Chen, J., Kim, C.-S., Song, F.: A Distributed Clustering Algorithm for Voronoi Cell-Based Large Scale Wireless Sensor Network. In: Proceedings of International Conference on Communications and Mobile Computing (CMC), Shenzhen, China, pp. 209–213 (2010)
9. West, D.B.: Introduction to Graph Theory, 2nd edn. Prentice Hall, New Jersey (2001)

A Trustworthiness Evaluation Method for Wireless Sensor Nodes Based on D-S Evidence Theory

Chenglin Miao, Liusheng Huang, Weijie Guo, and Hongli Xu

Dept. of Computer Sci. & Tech. / Suzhou Inst. for Advanced Study
University of Science & Technology of China
Hefei, Anhui 230027, P.R. China
{clmiao,ustcer86}@mail.ustc.edu.cn,
{lshuang,xuhongli}@ustc.edu.cn

Abstract. As many wireless sensor networks (WSNs) are deployed in complicated environment without good physical protection, the sensor nodes are more vulnerable to be affected by uncertain factors from inside or outside so that the sensed data always cannot reflect the real world situation well. Thus the trustworthiness of sensor nodes should be evaluated for revising the faulty ones in after-deployment maintenances. In this paper, we propose a trustworthiness evaluation method based on D-S evidence theory in data level for sensor nodes which can sense multi-dimensional data. Different dimensions of a sensor node are regarded as its different trustworthiness attributes in this method. For a single node, the trustworthiness of each attribute is evaluated firstly based on evidence theory, and then the lower and upper limits of trust degree for this node are calculated by fusing the evaluation results of different attributes. Moreover, in order to figure out whether regional uncertain factors exist or not, the trust degree of a local region is given by fusing the judgments of deployed sensor nodes according to the combination rules of evidence theory. Extensive experiments based on actual data samples are conducted to evaluate the performance of our method. The theoretical analysis and experimental results show that our method can give effective trustworthiness evaluation for one single sensor node or a local region. Also, robustness and stability of this method are verified in the experiments.

Keywords: wireless sensor nodes, multi-dimensional data, trustworthiness evaluation, D-S evidence theory.

1 Introduction

Wireless sensor networks (WSNs) have been widely used in many important application domains such as infrastructure protection, environment monitoring, and scientific exploring [1]. These systems should run reliably in the context of real world to collect data information and sustain for years. However, most of

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 163–174, 2013.

sensor nodes in these systems are implemented using cheap hardware components which are easily influenced by some uncertain environmental factors, such as electromagnetic interference or physical crash. Thus, they are error-prone and subject to component faults, performance degradations and even major system failures in real world deployments [2][3][4]. All of these matters may make the data collected by sensor nodes cannot reflect the real world situation accurately and the faulty data often mislead us to make wrong judgments.

Considering the importance of information assurance in WSNs, e.g., in a surveillance network [5], we should deliberate the trustworthiness of wireless sensor nodes before the readings of them are used for further judgments. In order to collect accurate data information, many schemas have been designed to be fault tolerant to some extent, such as removing nodes with faulty readings from a system with some redundancy or replacing them with good ones. These methods can significantly improve the whole system's performance and prolong the lifetime of the network at the same time. But before conducting such after-deployment maintenances (e.g., remove and replace), it is essential to investigate methods for evaluating the trust degree of sensor nodes in order to make sure which one is untrustworthy.

For the purpose of evaluating trustworthiness for sensor nodes, researchers have proposed many evaluation methods. For example, [15] presented a model which needed actions and interacting rules of evaluated individuals. But most of these methods are based on behaviors of individuals or communications between nodes, and the complex evaluation works need to be finished by sensor nodes themselves. However, the energy on one sensor node is so limited that it should try its best to save energy for regular data transmission. Especially for sensor node which is able to collect multi-dimensional data (e.g., sensing solar radiation, air temperature, air related humidity and so on at the same time) of the environment, they should save more energy as they have more data to manage. So a trustworthiness evaluation method which operates independently from sensor nodes is essential to this kind of applications.

Motivated by above problems, we propose a novel trustworthiness evaluation method based on D-S evidence theory [6][7] in data level for wireless sensor nodes with multi-dimensional data. This method can operate at a base station independently without the participation of sensor nodes. According to historical normal data and current readings of nodes, the trustworthiness can be effectively evaluated not only for a single sensor node but also for a local region with our method. Based on the evaluation results, we can find untrustworthy nodes and figure out whether some regional influence factors (such factors always interfere many sensor nodes in a region at the same time rather than influence one of them) exist or not.

The main contributions of this paper are summarized as follows:

- We propose a method to give the lower and upper limits of trust degree for one sensor node which can collect multi-dimensional data of the environment.
- Considering the impacts of some regional uncertain factors, we design a trustworthiness evaluation method for a local region based on the judgments of sensor nodes deployed in it.

- Extensive experiments are conducted on actual data sets which were collected by some meteorological sensor nodes and the effectiveness and stability of our method are validated also.

The rest of this paper is organized as follows. Section 2 summarizes the related works. Section 3 describes the details of trustworthiness evaluation method based on D-S evidence theory. Experimental results are given in section 4. And we conclude the paper in section 5.

2 Related Work

The trust problems have been focused for a long time by people [10]. It started in social sciences where trust among humans was studied. The effect of trust was also analyzed in economic transactions [11], and Marsh firstly introduced a computational model for trust [12]. Then [13] presented a notion to judge how trusted an internet seller can be in e-commerce. Recently, researchers have been concentrated their attentions on the concept of trust to increase security and reliability in Ad Hoc and sensor networks [14].

With regard to the trust management in WSNs, researchers have proposed many schemes based on sensor nodes behaviors [15] and the communications between nodes [8]. These trust management methods have advantages of peer-to-peer structure and no centers, but most of them need to operate on sensor nodes, which may lead to a large energy consumption for resource-constrained nodes. Moreover, some research analyzed uncertain factors which make the WSNs untrustworthy or unreliable, such as detecting the wireless sensor intrusion or finding the network failures [9], these works could help to improve the security of WSNs, but they could not make effective trustworthiness evaluation for sensor nodes or networks.

Marmol [16] gave a wide review of different trust models with an interface proposal and provided some pre-standardization recommendations at the same time. The best practices which are essential for developing a good trust management system in WSNs were listed by Lopez [17] and the author also made an analysis about the state of the art considering these practices. Both [16] and [17] gave an excellent summary, which introduced many profound viewpoints and pointed out an additional insight on the trustworthiness evaluation field. Additionally, some other protocols related to trust management in self-organization networks from different views were proposed in [18].

3 Details of Trustworthiness Evaluation Method

In this section, we describe the details of the method proposed to evaluate the trustworthiness of sensor nodes which collect multi-dimensional environmental data. The method operates in data level of sensor nodes based on evidence theory, and it mainly contains two parts: the trustworthiness evaluation for one sensor node and the evaluation for a local region. We first introduce D-S evidence theory before the description of the proposed method.

3.1 The D-S Evidence Theory

D-S evidence theory [6][7] is a generalization of the Bayesian theory of subject probability. It can distinguish between uncertainties and “unknown” when the prior probability is difficult to be given. Thus the theory has great flexibility, and is more appropriate for expert system than the probability theory. Also, this theory can reduce the uncertainty by allowing evidence to support several mutually exclusive conclusions.

Frame of Discernment: In the D-S theory, parameters such as events in probability theory are called propositions. For a recognition problem, we use $\Theta = \{\theta_1, \theta_2, \theta_3, \cdots, \theta_n\}$ to represent all possible outcomes as a *frame of discernment* [7]. A singleton proposition θ_i represents the lowest level of discernible information. All elements or element collections in set Θ are incompatible. The elements in 2^Θ,where 2^Θ denotes the power set of Θ, form all the propositions of interest. We use notation $\overline{A}$ to denote all singletons in Θ that are not included in A, where $A \subseteq \Theta$.

Basic Belief Assignment Function: Based on the frame of discernment Θ, a function $m : 2^\Theta \mapsto [0, 1]$ (defined on 2^Θ, and the range of its value is [0,1]) is called a basic belief assignment function if

$$\begin{cases} \sum\limits_{A_i \subseteq \Theta} m(A_i) = 1 \\ m(\emptyset) = 0 \end{cases} \tag{1}$$

where $m(A)$ reflects the reliability of A itself. As a result, $m(\emptyset) = 0$, where $\emptyset$ denotes the set of impossible event and $\sum\limits_{A_i \subseteq \Theta} m(A_i) = 1$, which shows the probability of total events is 1.

Belief Function: Based on the frame of discernment Θ, function $Bel : 2^\Theta \mapsto [0, 1]$, for any subset A of Θ, is called the belief function (i.e., lower limit function) if

$$\begin{cases} Bel(A) = \sum\limits_{B \subseteq A} m(B) \\ Bel(\Theta) = \sum\limits_{B \subseteq \Theta} m(B) = 1 \\ Bel(\emptyset) = m(\emptyset) = 0 \end{cases} \tag{2}$$

and the value of Bel represents the total reliability of A.

Plausibility Function: The same to Bel, if function $Pl : 2^\Theta \mapsto [0, 1]$, for any subset A of Θ, satisfies

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B) = 1 - \sum_{A \cap B = \emptyset} m(B) \tag{3}$$

also,

$$Pl(A) = 1 - m(\overline{A}) \tag{4}$$

then Pl is the plausibility function, also called upper limit function or irrefutable function.

In D-S evidence theory, there is $0 \leq Bel(A) \leq Pl(A) \leq 1$, where $Bel(A)$ and $Pl(A)$ are called lower limit uncertainty value and upper limit uncertainty value of A. $Pl(A) - Bel(A)$ is the measurement of neither trust A nor trust $\overline{A}$ (the measurement of "unknown"). The confidence interval of D-S evidence theory is shown in Fig. 1:

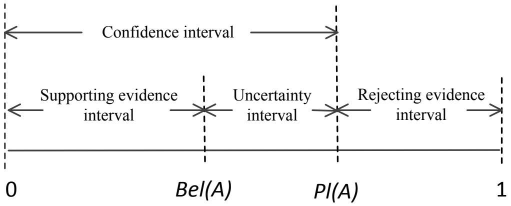


Fig. 1. The confidence interval of D-S evidence theory

Dempsters Rule of Combination: Let $m_1, m_2, \cdots, m_n$ be basic belief assignment functions on the same frame of Θ. Intuitively, $m_i(U)$ describes the extent to which the evidence supports U, where $U \in 2^{\Theta}$, $i = 1, 2, \cdots, n$. Then the rule of combination by D-S evidence theory is:

$$m(A) = \begin{cases} \frac{\sum\limits_{\cap_i A_j = A} \prod\limits_{1 \leq i \leq n} m_i(A_j)}{1 - \sum\limits_{\cap_i A_j = \emptyset} \prod\limits_{1 \leq i \leq n} m_i(A_j)}, & if\ A \neq \emptyset \\ 0, & if\ A = \emptyset \end{cases} \tag{5}$$

where A, $A_j \in 2^{\Theta}$.

3.2 Trustworthiness Evaluation for One Sensor Node

As sensor nodes in WSNs are always deployed in the complex environment, they may easily suffer influences from uncertain factors of outside or inside the WSNs. In this situation, the data sensed by these nodes may not reflect the real world environment well. So we should do some after-deployment maintenances for the problematic nodes. But before this action, we ought to evaluate the trustworthiness of sensor nodes to find which one need to be replaced or removed. Here, we consider the application scenes where one sensor node can sense multi-dimensional data of the environment.

In this paper, we define different dimensions of one sensor node as its different *trustworthiness attributes* (i.e., n dimensions correspond to n trustworthiness attributes). If a node works well or we trust it, the sensed data in each dimension of the node should be able to reflect the real world environment accurately. Thus, we can obtain current trust degree of a node by evaluating the current readings in each dimension of it.

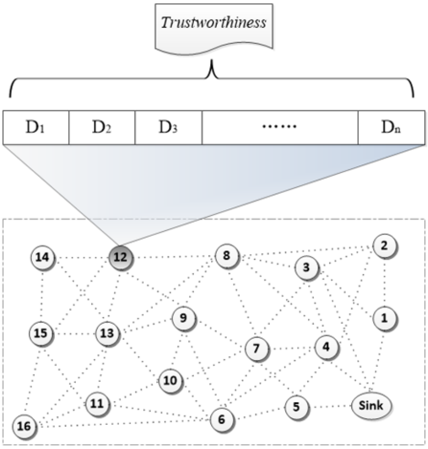


Fig. 2. The trustworthiness evaluation of one sensor node

As shown in Fig. 2, we suppose that lots of sensor nodes are deployed in an application system of WSNs and each node can collect n-dimensional data. Here we let $D_1, D_2, D_3, \cdots, D_n$ denote n dimensions respectively, such as node 12 (i.e., the gray one) in Fig. 2. If we doubt the authenticity of node 12 according to current readings, we should evaluate the trustworthiness of it. And the evaluation can be considered as the measurement of the proposition P : $\{N\ is\ trustworthy\}$, where N is the node we need to evaluate. Based on the *trustworthiness attributes* we defined above, the proposition P can be divided into $P = \{P_1, P_2, \cdots, P_n\}$, where $P_i : \{D_i\ is\ trustworthy\}$ and $i = 1, 2, \cdots, n$. Additionally, our paper focuses attention on the applications in which different dimensions of one sensor node correlate with each other in data level to some extent (e.g., solar radiation, air temperature and air humidity), and we can use one dimension to evaluate other dimensions with this correlation which can be learned from historical normal data. So according to (2), the belief function of P for node N based on D_h $(h = 1, 2, \cdots, n)$ is:

$$\begin{aligned} Bel_h(P) &= Bel_h(P_1, P_2, \cdots, P_n) \\ &= \sum_{1 \le i \le n} m_h(P_i) + \sum_{1 \le i,j \le n, i \ne j} m_h(P_i, P_j) \\ &+ \sum_{1 \le i,j,k \le n, i \ne j, i \ne k, j \ne k} m_h(P_i, P_j, P_k) \\ &+ \cdots + m_h(P_1, P_2, \cdots, P_n) \end{aligned} \tag{6}$$

According to (3) and (4), the plausibility function of P for node N based on D_h is:

$$\begin{aligned} Pl_h(P) &= 1 - Bel_h(\overline{P}) \\ &= 1 - Bel_h(\{\overline{P_1, P_2, \cdots, P_n}\}) \\ &= 1 - Bel_h(\{\overline{P_1}, \overline{P_2}, \cdots, \overline{P_n}\}) \end{aligned} \tag{7}$$

where

$$\begin{aligned} &Bel_h(\{\overline{P_1}, \overline{P_2}, \cdots, \overline{P_n}\}) \\ &\quad = \sum_{1 \le i \le n} m_h(\overline{P_i}) + \sum_{1 \le i,j \le n, i \ne j} m_h(\overline{P_i}, \overline{P_j}) \\ &\quad + \sum_{1 \le i,j,k \le n, i \ne j, i \ne k, j \ne k} m_h(\overline{P_i}, \overline{P_j}, \overline{P_k}) \\ &\quad + \cdots + m_h(\overline{P_1}, \overline{P_2}, \cdots, \overline{P_n}) \end{aligned} \tag{8}$$

In the procedure of trustworthiness evaluation, we use $[\overset{\vee}{Tr}, \overset{\wedge}{Tr}]$ to indicate the final result of trustworthiness evaluation for one node, where $\overset{\vee}{Tr}$ is the lower limit of trust degree for one sensor node and $\overset{\wedge}{Tr}$ is the upper limit. In this paper, we set the trustworthiness within the interval of $[0, 1]$. If the values of $\overset{\vee}{Tr}$ and $\overset{\wedge}{Tr}$ are near to 1, it means the evaluated sensor node is trustworthy. If the values of $\overset{\vee}{Tr}$ and $\overset{\wedge}{Tr}$ are near to 0, the evaluated node is untrustworthy. The higher the values of $\overset{\vee}{Tr}$ and $\overset{\wedge}{Tr}$, the more trustworthy the evaluated sensor node. After acquiring the trustworthiness evaluation based on each dimension, we calculate $\overset{\vee}{Tr}$ and $\overset{\wedge}{Tr}$ as follows:

$$\overset{\vee}{Tr} = \frac{1}{n} \sum_{1 \le h \le n} Bel_h(P) \tag{9}$$

$$\overset{\wedge}{Tr} = \frac{1}{n} \sum_{1 \le h \le n} Pl_h(P) \tag{10}$$

And the values of basic belief assignment functions we used above can also be learned from the historical normal data which were collected by these sensor nodes deployed in the same environment. After acquiring the lower and upper trustworthiness limits of one node using this method, we should deliberate the results based on the demands of the application system for the environment and then take further measures (e.g., replace and remove) for the node whose trustworthiness limits cannot meet our needs (e.g., $\overset{\vee}{Tr}$ and $\overset{\wedge}{Tr}$ are all lower than 0.5).

3.3 Trustworthiness Evaluation for a Local Region

In the complex environment where sensor nodes are deployed, some regional uncertain factors (e.g., electromagnetic wave) always interfere many sensor nodes in a local region rather than one of them. Because the main reasons for low trustworthiness of sensor nodes come from the outside environment in this situation, we should consider to take measures for the whole region rather than replacing or removing some nodes easily in the after-deployment maintenances(e.g., take measures to eliminate the interference of electromagnetic in this region). But before taking such measures, we should figure out whether the regional uncertain factors exist or not by evaluating the trustworthiness of this region with the evidences of sensor nodes deployed in it.

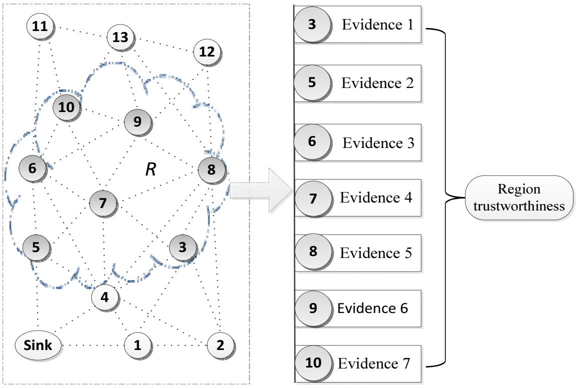


Fig. 3. The trustworthiness evaluation of a local region

Similar to the trustworthiness evaluation for one sensor node, we see different nodes in a local region as different *trustworthiness attributes* for this region. If most of sensor nodes in a region have crisis of trustworthiness or we doubt that there are uncertain factors in the region, we should evaluate the trust degree of it. As shown in Fig. 3, we doubt that there are some interference factors to nodes in the cloud shape region R, so we select some nodes in it (i.e., node 3, 5, 6, 7, 8, 9, 10 in the figure) as the *trustworthiness attributes* of R. Each node selected here can make a judgment about the trustworthiness of this region and they are different evidences (as shown in the right part of Fig. 3) for the proposition $P' : \{R\ \textit{is trustworthy}\}$. Through fusing these evidences based on D-S evidence theory, we can get the trust degree of region R.

Considering the proposition P', we set the discernment as $W = \{w, \neg w\}$, where w and $\neg w$ represent two trust states, namely credible and incredible. 2^W is $\{\emptyset, \{w\}, \{\neg w\}, \{w, \neg w\}\}$, in which $\emptyset$, $\{w\}$, $\{\neg w\}$ and $\{w, \neg w\}$ represent respectively the empty set, the propositions of region's 'Trust', 'Distrust' and 'Uncertain'. Also, we assume that there are k nodes $(n_1, n_2, \cdots, n_k)$ are selected to evaluate the trustworthiness of one local region and suppose m_i $(i = 1, 2, \cdots, k)$ are the basic belief assignments over the frame of discernment W, and these assignments can be obtained from part 3.2 of this section (we set $m_i(\{w\})$ as $\overset{\vee}{Tr}_i$, $m_i(\{\neg w\})$ as $1 - \overset{\wedge}{Tr}_i$ and $m_i(\{w, \neg w\})$ as $\overset{\wedge}{Tr}_i - \overset{\vee}{Tr}_i$). Intuitively, $\{w\}$, $\{\neg w\}$ and $\{w, \neg w\}$ respectively denote the support level of trust evidence to the 'Trust', 'Distrust', and 'Uncertain' grade. According to (5), we have

$$m(\{w\}) = \frac{\sum\limits_{\cap_i A_j = w} \prod\limits_{1 \leq i \leq k} m_i(A_j)}{1 - \sum\limits_{\cap_i A_j = \emptyset} \prod\limits_{1 \leq i \leq k} m_i(A_j)} \tag{11}$$

where $A_j \in 2^W$.

Obviously, $m(\{w\})$denotes the region's trust degree and if we find the result is very low, we can infer that there may be some regional uncertain factors which can interfere the normal function of many sensor nodes in this region. Thus, we should take further measures to eliminate these uncertain factors in after-deployment maintenances. In this procedure, we may find the judgment of one sensor node conflict with others. Here we eliminate the node and do the evaluation using others.

4 Experiments and Results

In order to verify the effectiveness of our method, extensive experiments based on a data set collected by some meteorological sensor nodes are conducted in this section. We use the real world climate data from U.S. National Climatic Data Center [19]. Each node here can collect 3-dimensional data and they are daily readings of Maximum temperature ($^\circ F$), Minimum temperature ($^\circ F$), and Precipitation ($inch$). The data set covers 1064 stations for nearly the last 100 years.

In the experiment of trustworthiness evaluation for one sensor node, we select dozens of nodes as samples and evaluate their trustworthiness according to the data they collected most recently. As these sensor nodes worked normally and the data used here are all reliable, we randomly modify the data in some dimensions of them to abnormal states (these abnormal states are often caused by some uncertain factors) in order to simulate the interference of uncertain factors. The basic belief assignment functions used here are obtained by training historical normal data. Through the experiment, we find our method works well for evaluating the trustworthiness of these "interfered" nodes. For easy of illustration, we select five nodes (N_1, N_2, N_3, N_4, N_5) in the samples to show the results. As shown in Table 1, we give the lower and upper trustworthiness limits ($\overset{\vee}{Tr}$ and $\overset{\wedge}{Tr}$) of these nodes in different states: normal state (S_1), one dimension modified (S_2), two dimensions modified (S_3) and three dimensions modified(S_4).

From Table 1 we can see that when the data is not interfered (state S_1), the upper limit of trustworthiness for node N_3 is more than 90% and others are also near to this value (the reason why it cannot be 100% is that these data have small fluctuations compared with historical normal data actually although they are all reliable). But when one dimension or two dimensions are interfered (state S_2 and S_3), the upper limits become very low: in state S_2 the value is 50%$\sim$60% and in state S_3 the value is 20%$\sim$30%. When all the dimensions are modified randomly to abnormal state, the trustworthiness of these nodes will be near to 0 and we cannot trust them anymore so that we must take measures in the after-deployment maintenances.

Table 1. Evaluation Results of N_1, N_2, N_3, N_4, N_5

Nodes	S_1		S_2		S_3		S_4	
	$\overset{\vee}{Tr}$	$\overset{\wedge}{Tr}$	$\overset{\vee}{Tr}$	$\overset{\wedge}{Tr}$	$\overset{\vee}{Tr}$	$\overset{\wedge}{Tr}$	$\overset{\vee}{Tr}$	$\overset{\wedge}{Tr}$
N_1	0.71135	0.89372	0.47318	0.64221	0.08433	0.23621	0.00135	0.03674
N_2	0.69837	0.87401	0.45515	0.53365	0.18132	0.30124	0.00074	0.00961
N_3	0.73361	0.91620	0.47324	0.60211	0.08711	0.24689	0.00083	0.01503
N_4	0.68896	0.85289	0.41303	0.65379	0.06545	0.19187	0.00112	0.04761
N_5	0.70901	0.87341	0.38402	0.54434	0.10367	0.26173	0.00094	0.02433

The variation trends of $\overset{\vee}{Tr}$ and $\overset{\wedge}{Tr}$ about the five nodes in different states are shown in Fig. 4 (S_1, S_2, S_3, S_4 are the states we list above and the dots of color are the values of trust degree limits for different nodes). The results in the figure show that the more dimensions are "interfered" to abnormal state, the lower trustworthiness of the node is, especially when all the dimensions are interfered irregularly, the trustworthiness of these nodes are so low that we should remove or replace them. In each state, as shown in the figures, the values of trust degree limits for the five nodes are relatively centralized although they are different with each other. And these trends illustrate that our method can work effectively and stably.

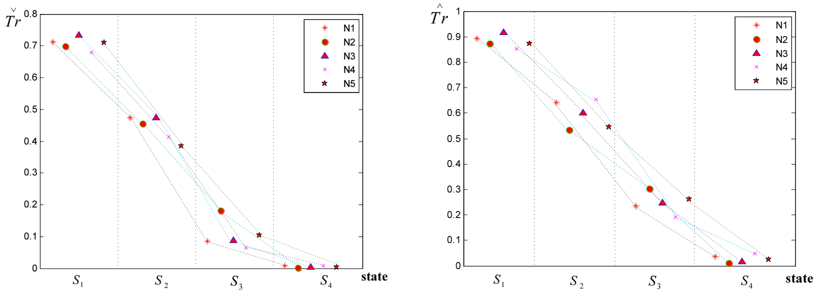


Fig. 4. The lower and upper trustworthiness limits in different states

Moreover, for further verifying the effectiveness and stability of our method, we operate our method on 500 sensor nodes with some "interfered" data and calculate the average and variance of upper trustworthiness limits ($\overset{\wedge}{Tr}$) in different states. As shown in Table 2, the more dimensions are modified to abnormal state, the lower average value of $\overset{\wedge}{Tr}$ is. Additionally, the values of variance about $\overset{\wedge}{Tr}$ are always very small in different states. And these results show our method is very robust and stable.

In the experiment for a local region, we select 10 sensor nodes whose data are close with each other and we assume they are deployed in a local region. For the purpose of simulating the interferences of regional uncertain factors to sensor nodes in this region, we modify the data of these nodes to varying

Table 2. the Average and Variance of $\overset{\wedge}{Tr}$ for 500 Sensor Nodes

State of Nodes	$\mathbf{E}(\overset{\wedge}{Tr})$	$\mathbf{D}(\overset{\wedge}{Tr})$
S_1	0.89569	0.0010235
S_2	0.59364	0.0016842
S_3	0.23331	0.0009471
S_4	0.02107	0.0002455

degrees and evaluate the trustworthiness of this region with the data which are "interfered" by us. Through the simulation we find that the trustworthiness value of this region is nearly 90% when these adjacent nodes are normal (the data is not interfered by uncertain factors) and the value becomes lower when there are more "interfered" nodes (especially more dimensions of these nodes are modified). When most dimensions of these nodes in the region are "interfered", we find that the trustworthiness value is near to 0. In this case, we can infer that the interference may not be caused by these nodes themselves but some reasons in the region because most of these adjacent nodes do not trust the region very well and they show low trustworthiness. Thus, measures should be taken to find the sources of this interference and then eliminate them. However, we cannot point out which kind of interference it is although we can be aware of the regional uncertain factors exist, so we will do more researches in the next step to find valid methods about this problem.

5 Conclusion

In this paper, we present a trustworthiness evaluation method for wireless sensor nodes which can collect multi-dimensional data. This method operates in data level based on D-S evidence theory and the historical normal data of these nodes. Through the evaluation, the trust degree can be obtained not only for one single sensor node but also for a local region and the results can provide good advices for after-deployment maintenances of WSNs. Extensive experiments based on actual data sample are also conducted in this article, and experimental results show that our method can effectively evaluate the trustworthiness of one single node or a local region. Besides, we verify the robustness and stability of this method. Although the trustworthiness can be evaluated effectively, pointing out which kind of regional uncertain factors exist in a local region is still difficult in our method. So we will do more research on this problem in the future to improve this method.

Acknowledgment. This paper is supported by the National Grand Fundamental Research 973 Program of China under Grant No. 2011CB302905, National Science and Technology Major Project under Grant No. 2011ZX03005-004-04 and 2012ZX03005009, Special Project on IoT of China NDRC (2012-2766), the National Science Foundation of China under Grant No. 61170058, 61272133 and 61228207, Special Project on Industry Technology of Suzhou (No. ZXG201041), and Guangdong Province and CAS Comprehensive Strategic Cooperation Projects under Grant No. 2012B090400013.

References

1. Tolle, G., Polastre, J., Szewczyk, R., Turner, N., Tu, K., Burgess, S., Gay, D.: A Macroscope in the Redwoods. In: SenSys. 2005 (2005)
2. Xing, K., Liu, F., Cheng, X., Du, D.H.C.: Real-time detection of clone attacks in wireless sensor networks. In: Proceedings of IEEE ICDCS, Beijing, China (2008)
3. Magistretti, E., Gurewitz, O., Knightly, E.: Inferring and mitigating a links hindering transmissions in managed 802.11 wireless networks. In: Proceedings of ACM MobiCom, Chicago, Illinois, USA (2010)
4. Wang, X., Fu, L., Hu, C.: Multicast performance with hierarchical cooperation. IEEE/ACM Transactions on Networking 20, 917–930 (2011)
5. He, T., et al.: VigilNet. An Integrated Sensor Network System for Energy-Efficient Surveillance. ACM Transactions on Sensor Networks 2, 1–38 (2006)
6. Dempster: Upper and lower probabilities induced by multivalued mapping. Annals of Mathematical Statistics 38(2), 325–339 (1967)
7. Shafer, Glenn: A Mathematical Theory of Evidence. Princeton University Press (1976)
8. Momani, M., Challa, S., Alhmouz, R.: Bayesian Fusion Algorithm for Inferring Trust in Wireless Sensor Networks. Journal of Networks 5(7) (July 2010)
9. Ma, Q., Liu, K., Miao, X., Liu, Y.: Sherlock is Around: Detecting Network Failures with Local Evidence Fusion. In: INFOCOM (2012)
10. Momani, M., Aboura, K., Challa, S.: RBATMWSN. Recursive Bayesian Approach to Trust Management in Wireless Sensor Networks. In: The Third International Conference on Intelligent Sensors, Sensor Networks and Information, Melbourne, Australia (2007)
11. Ba, S., Pavlou, P.A.: Evidence of the effect of trust building technology in electronic markets: price premiums and buyer behavior. MIS Quarterly 26 (2002)
12. Marsh, S.: Formalising Trust as a Computational Concept. In: Departmet of Computer Science and Mathematics. PhD, University of Stirling, p. 184 (1994)
13. McKnight, D.H., Chervany, N.L.: Conceptualizing Trust: A Typology and Ecommerce Customer Relationships Model. Presented at Proceedings of the 34th Hawaii International Conference on System Sciences (2001)
14. Srinivasan, A., Teitelbaum, J., Wu, J.: DRBTS. Distributed Reputation-based Beacon Trust System. In: 2nd IEEE International Symposium on Dependable, Autonomic and Secure Computing (2006)
15. Li, N.H., Mitchell, J.C., Winsborough, W.H.: Beyond Proof-of-Compliance: Security Analysis in Trust Management. J. ACM 52, 474–514 (2005)
16. Marmol, F.G., Perez, G.M.: Towards Pre-standardization of Trust and Reputation Models for Distributed and Heterogeneous Systems. Comput. Stand. Interfaces 32, 185–196 (2010)
17. Lopez, J., Roman, R., Agudo, I.: Trust Management Systems for Wireless Sensor Networks: Best Practices. Comput. Commun. 33, 1086–1093 (2010)
18. Li, J.L., Gu, L.Z., Yang, Y.X.: A New Trust Management Model for P2P Networks with Time Self-Decay and Subjective Expect. J. Electron. Inf. Technol. 31, 2786–2790 (2009)
19. http://www.ncdc.noaa.gov/oa/climate/research/ushcn/daily.html

UCOR: An Unequally Clustering-Based Hierarchical Opportunistic Routing Protocol for WSNs

Ziwei Liu[1,2], Chuanbo Wei[1], Yang Ma[1], Hui Li[2],
Xiaoguang Niu[1,*], and Lina Wang[1]

[1] Computer School, Wuhan University
[2] Institute of Seismology, China Earthquake Administration
{lzw,lihui}@eqhb.gov.cn,
{wchuangbo,mayang,xgniu,lnwang}@whu.edu.cn

Abstract. Recently, large amounts of researches have demonstrated that dynamic changes of topological structure as well as unbalance of data traffic/energy consumption in wireless sensor networks have made routing protocol expansibility challenging work. According to the phenomenon that large-scale wireless sensor network will shorten longevity due to funneling effect in energy consumption; we proposed Unequally Clustering-based Hierarchical Opportunistic Routing (UCOR) protocol to reduce redundant network traffic. The protocol can further increase packets delivery ratio, decrease network delay and achieve more balanced and lower energy consumption. Experiment results show that UCOR protocol prolongs network longevity by about 20% to support for larger scale network, compared with HEED, EEUC, etc.

Keywords: wireless sensor network, funneling effect, unequally clustering, equal energy consumption, opportunistic routing.

1 Introduction

With the development of wireless communication, wireless sensor networks (WSNs) have great potential in a wide variety of applications [1]. These systems have typically been characterized as being composed of a large number of sensor nodes with extreme resource constraints, which are deployed over a vast terrain to deliver data reports over multi-hop wireless paths to the sink. Unbalanced of energy consumption in sensor network has generally aroused researchers' attention [2].When sensor nodes send collected data to sink node, traffic in the network is nonuniformly distributed, decreasing from the sink node like a funnel [3]. As shown in Fig.1, when it's closer to sink node, network traffic becomes heavier. Correspondingly, sharp area in funnel has severe package collision and network congestion, which will directly result in exhaustion of energy of partial nodes, decrease in data delivery ratio, false detection or loss of important events and other problems in practical application.

[*] Corresponding author.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 175–185, 2013.

Simulation in [3] shows that there is up to 90% energy unused in wireless sensor networks, which follows random distribution, due to the funneling effects of energy consumption.

In order to decrease disequilibrium of energy consumption in WSNs, researchers have proposed various solutions, such as energy-aware routing protocol, clustering network design [4], topological control based on node unequal distribution [5] and channel contention/distribution mechanism [3]. However, those methods are not able to eliminate funneling effect thoroughly.

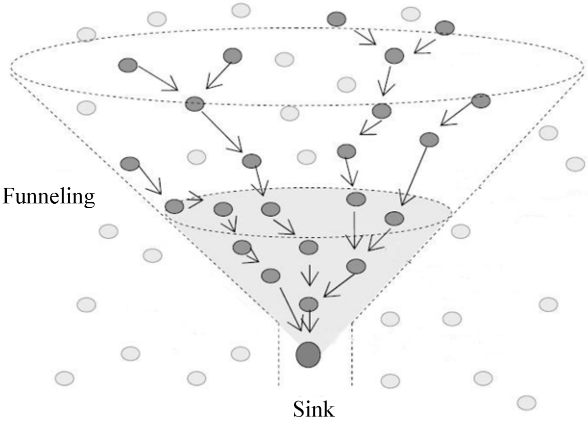


Fig. 1. The funneling effect in WSNs [3]

In the paper, Unequally Clustering-based Hierarchical Opportunistic Routing (UCOR) protocol is proposed to reduce redundant network traffic, improve network throughput in top region of funnel, and balance energy consumption. In UCOR protocol, network is unequally clustered based on characteristics of network traffic and energy consumption distribution, appropriate data aggregation methods are designed in different regions and opportunistic routing methods are adopted in various layers of network. Simulation experiments demonstrated that UCOR protocol reduces funneling effect of energy consumption in large-scale wireless sensor network, which makes more balanced and lower energy consumption, prolongs network longevity and supports for larger scale network.

The remainder of this paper is organized as follows. Section 2 discusses the related works. Our proposed UCOR protocol is presented in details in Section 3. Finally, we evaluate the performance of UCOR protocol in Section 4, and make a conclusion in Section 5.

2 Related Works

Researchers proposed various solutions to unbalanced energy consumption problem in wireless sensor network in different directions.

According to network load and channel contention conditions in each region, Funneling-MAC [3] designs mixed channel competition/distribution mechanism: applying TDMA method around sink node, in order to increase network throughput rate by reducing the packet collision, saving energy consumption in top region of funnel. Siva et al. proposed a centralized routing protocol called Base Station Controlled Dynamic Clustering Protocol (BCDCP) [12], which distributes the energy dissipation evenly among all sensor nodes to improve network lifetime and average energy savings. This method assumes that the properties of a given sensor network model are a fixed base station, sensor nodes with energy constraints, nodes equipped with power control capabilities, and stationary nodes. However, with the increasing scale of network, maintenance cost of TDMA will rise rapidly when application areas expand, which will counteract decrease of network load, further aggravating network funneling effect.

Soro et al. proposed unequally clustering to solve funneling effect by dividing network into concentric circles in two layers, surrounding the sink node [11]. In HEED protocol [8], network is clustered. Cluster header collects data within the cluster and sends sensed data directly to the sink node by changing transmit power optionally in order to avoid funneling effect in energy consumption. In EEUC protocol [9], nodes not simply choose cluster header which is nearest to them, but take into consideration the distance between cluster header candidates and the sink node. In order to balance load of cluster header, clusters are divided into different size. Wang et al. [13] and Nurhayati et al. [14] proposed unequal-cluster routing protocols with unequal transmission power, in which nodes can select the proper power level to communicate with the sink node based on the received signal strength from the sink node. Liu et al. proposed a non-uniform clustering algorithm to balance the energy consumption and prolong the lifetime of the network [15]. Those protocols has a shortcoming that network communication capacity is decreased because large transmitting radius makes nodes in other region of the network are not able to transmit simultaneously.

3 UCOR: Unequally Clustering-Based Hierarchical Opportunistic Routing Protocol

In this section, we present the basic features of UCOR protocol, deferring its analysis for the next section.

3.1 Architecture

In UCOR protocol, network is divided into two parts, a region with heavy load and a set of cluster ring containing many clusters that become larger when their distance from the sink node increases, as shown in Fig. 2.

UCOR protocol can balance energy consumption and prolong network longevity by two following methods:

1) Unequally clustering algorithm. Network is unequally clustered, cluster radius increasing with its distance from the sink node. In order to decrease traffic to the top region of funnel around the sink node, take additional energy consumption in region far from the sink as expense to eliminate redundant data.

2) Hierarchical opportunistic forwarding algorithm. Adopting hierarchical packet transmitting mechanism based on opportunism forwarding, which can accomplish data collection within cluster and data transmission between cluster header and the sink node, improve data transmission reliability and decrease hop counts to reduce traffic at the top of funnel. Because traffic around the sink node sharply reduces and opportunistic data forwarding mechanism relieve packets collision, further increasing network throughput rate, UCOR primarily remits funneling effect which is common in wireless sensor network, and prolong network longevity and support for larger scale network.

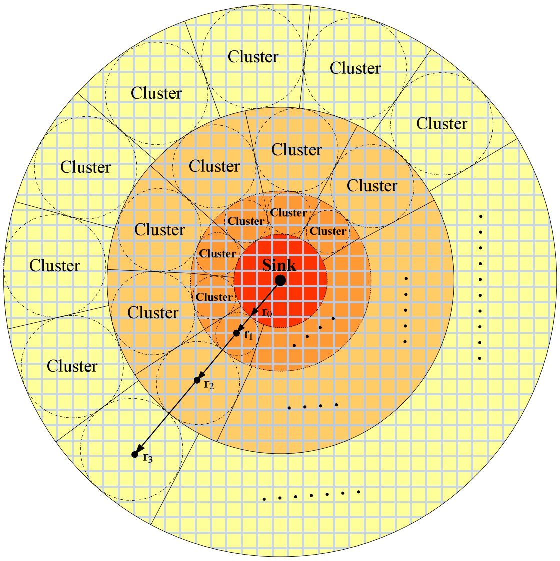


Fig. 2. Unequally-clustered mechanism in wireless sensor networks

3.2 Unequally Clustering Process

Unequally clustering algorithm in UCOR protocol is a distributed contention algorithm based on comparison of dump energy in each node. It's a static unequally clustering algorithm, network is clustered in initialization phase after network is deployed and makes each sensor node belong to a certain cluster within entire network life cycle. In the initialization state, the sink node firstly determine cluster radius in different regions based on distance to itself according to network scale and then inform corresponding nodes in the network that they are cluster header candidates. As shown in Fig.2, regions where it's in a circular area with r0 distance from the sink node generally have high load. Nodes in regions with r1, r2, r3 distance from the sink node are selected to be initial cluster header candidates.

Correspondingly, cluster radiuses are $\lceil r_1 - r_0 \rceil$, $\lceil r_2 - 2r_1 + r_0 \rceil$ and $\lceil r_3 - 2r_2 + 2r_1 - r_0 \rceil$ .The sink node will aggregate sample data that are directly transmitted from nodes in regions with heavy load. Initial cluster header candidates participate in election by broadcasting R-hop cluster header election packet based on their cluster radius R. After election, the number of cluster headers in circular regions with distance r_1, r_2, r_3 from the sink node (i.e. NC_{r1}, NC_{r2}, NC_3) are respectively obtained as following:

$$NC_{r1} = \frac{2\pi r_1}{\lceil r_1 - r_0 \rceil}, NC_{r2} = \frac{2\pi r_2}{\lceil r_2 - 2r_1 + r_0 \rceil}, NC_{r3} = \frac{2\pi r_3}{\lceil r_3 - 2r_2 + 2r_1 - r_0 \rceil}$$

After the election for cluster headers, successful cluster headers will broadcast R-hop BEACON packets. Normal nodes choose a cluster header which costs them least energy for communication (receiving most BEACON packets from that cluster header) in the cluster, based on distance from the sink node to themselves, and send BEACON responding packets to inform that cluster header. Then, network centered with the sink node forms many cluster rings with different radius on circular rings at the distance of r_1, r_2, r_3 from the sink node.

Cluster header is the most important node in clustered sensor network. In order to relieve energy consumption of cluster header, in UCOR protocol, cluster headers are only responsible for maintenance of nodes within the belonged cluster and work on collection, aggregation and transmission of sample data from them. In order to balance energy consumption of nodes in every cluster, cluster header will decide whether to change the cluster header based on dump energy in each node within the cluster at the beginning of data collection period. According to clustering algorithm in UCOR protocol specifically, when dump energy E_{ch} of cluster header is γ lower than the highest dump energy E_{mmax} of other nodes, corresponding to the node n_a in the cluster, cluster header will appoint node n_a to be the new cluster header. In the paper, γ is a specific threshold value and $\gamma = E_{mmax} \times 20\%$. Cluster members record their operations to compute their own dump energy in time [6], and send the dump energy information within sample data packets to the cluster header by a piggybacking way during every sampling period, which doesn't need additional communication overhead.

3.3 Hierarchical Opportunistic Forwarding Process

In order to make sensor network collect data more effectively, UCOR protocol divides communication of data packets into two levels, in-cluster sample data collection and aggregated data transmission between cluster header and the sink node, according to traffic distribution characteristics in unequal clusters.

In the process of data collection communication in the cluster, cluster members transmit sample data packets to their cluster header using COR [7]. The opportunistic forwarding process of aggregated data packets from cluster to cluster is described as following: 1) Cluster header aggregates sample data collected from its cluster members; 2) Aggregated data packet is transmitted to an adjacent cluster which is closer to the sink node through a common forwarding node; 3) Forwarding node in the next cluster receives the aggregated data packet and forwards it to next cluster closer to sink node till it arrives at sink. Thus, forwarding nodes in each cluster constitute a routing which links every two adjacent clusters. Since sensor network node density is high, the number of available node candidates for opportunistic data forwarding hugely exceeds the minimal number of node candidates which are essential to ensure transmission reliability. In order to improve high efficiency of energy consumption, there are only a few nodes that are selected to be forwarding nodes in each cluster.

Fig.3 shows opportunistic aggregated data forwarding routing from the cluster header in the 4th cluster ring to the sink node. When dump energy of forwarding node in any cluster is lower than that of other node which is as far as current forwarding node from the former forwarding node by a specified threshold, cluster will change current forwarding node and set up new forwarding routing. As a result, communications between clusters achieve balanced energy consumption through localized maintenance in UCOR protocol.

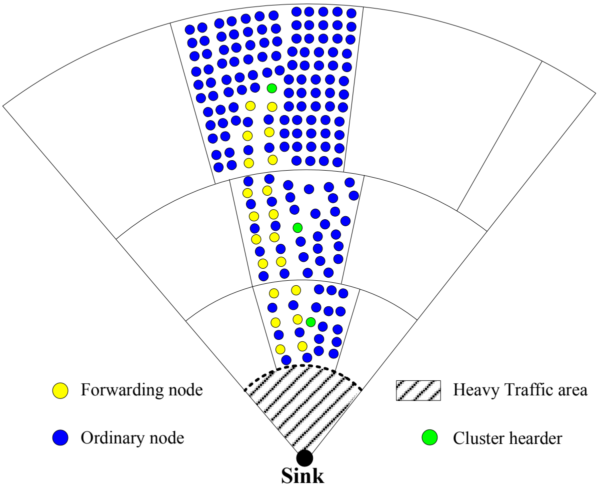


Fig. 3. Diagrammatic map of opportunistic data transmission from cluster to cluster in UCOR protocol

4 Performance Analysis

In this section, we analyze the performance characteristics of UCOR by comparing its energy equity and efficiency with two representative cluster-based routing schemes.

UCOR protocol is simulated in C language. Firstly, we take researches on energy efficiency of UCOR protocol with different parameters. To simplify the experiment, we assume an ideal MAC protocol and ignore data loss in wireless channel contention in the simulation. In addition, we assume that energy consumption of nodes of same kind in the same cluster ring is balanced. In the experiment, we calculate total energy consumption of receiving data, aggregating data, transmitting data and eavesdropping idle channel of nodes in the same region. In order to prove the high efficiency of UCOR protocol, we make a comparison between UCOR protocol and other two multi-hop routing protocols based on clusters: HEED-M [8] protocol and EEUC [9] protocol. Table 1 shows the parameters in experiment and those which are related to energy consumption model [10].

Table 1. Basic parameters configuration in simulation experiment

Parameter	Value
Field size	A round area with radius of 1000 meters, centered with sink node
Valid communication radius of node in single-path routing protocol (R)	100 meters
Communication bandwidth(*B*)	19.2 kilobit/s
Node number(*N*)	5024
Data package size(L_{pkt})	50 bytes
Data aggregation compression ratio(R_{df})	1/8
Initial energy of node(E_{total})	2000mAH
Energy consumption in Tx of RF (E_{send})	21.4mA（10dBm）
Energy consumption in Rx of RF (E_{rev})	7.5mA
Data aggregation energy consumption (E_{df})	6.9mA

We assume that data delivery ratio of node link is inversely proportion to the link length, and there is a random Gaussian offset with expectation value 0.1. In the simulation, sing-path routing's threshold value of data acceptance rate is 0.9 when it chooses a link (Table 1 shows that data acceptance of 100-meter link is 90%). The least applicable link data acceptance rate of opportunistic forwarding mechanism in UCOR protocol is 0.2. The experiment involved UCOR protocol in two conditions, respectively dividing the 10-hop network into 2 unequal cluster rings and 3 unequal cluster rings (UCOR-2, UCOR-3). In the two UCOR protocols, the radius of the clusters will choose an optimal value by repeated experiments, in which the network sampling period T is 15 minutes. In EEUC protocol, the ratio T of cluster header candidates to all the nodes in the network is 0.4, maximal cluster radius is 90 meters, the cluster radius range control parameter c is 0.5 and direct communication distance threshold TD_MAX is 140 meters. All the routing protocols in the experiments employs channel contention MAC protocol based on concentrated work/sleep, in which the duty cycle is 11.5%.

The stability of clustering topological structure has great influence on the network data transmission efficiency. We choose randomly 10 sampling periods in the experiments, and calculate the number of network clusters in each period. The result is shown in Fig.4. From the figure, we can find that the cluster number of UCOR-2 and UCOR-3 is far less than that of EEUC and HEED-M, and the cluster number remains the same in each period. That is because UCOR employs a static clustering method with multi-hop cluster radius and communication between nodes employs an opportunistic data forwarding mechanism with multi-candidate participating in the data forwarding. The data forwarding mechanism can loosen the limit of cluster radius, which makes cluster number in the network less and stable.

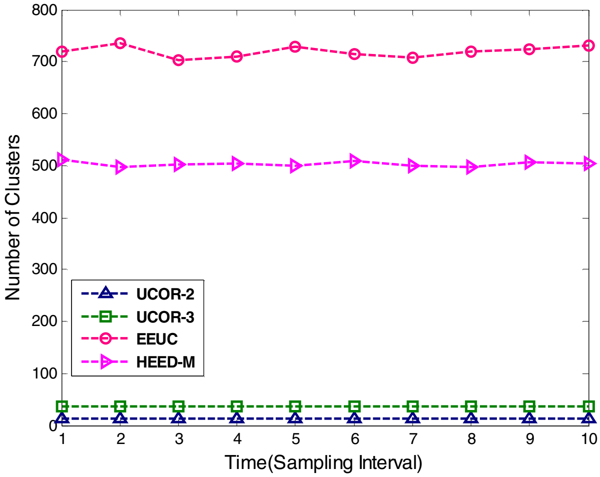


Fig. 4. Change of network cluster number

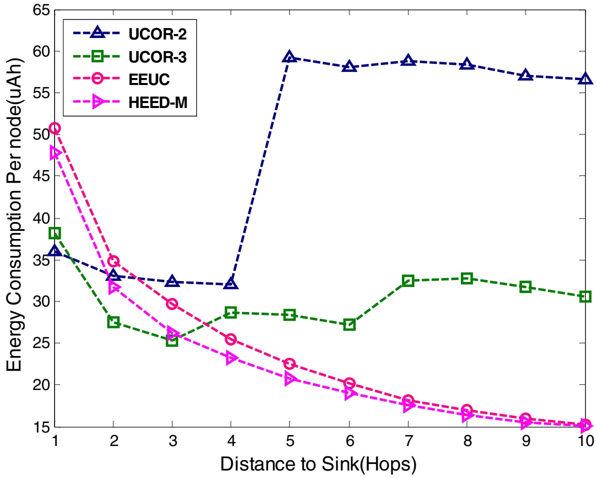


Fig. 5. Average energy consumption in unit period in different regions

Fig.5 and Fig.6 show the average energy consumption of nodes in each circular region which has different distance from the sink node. As shown in Fig.5, there is obvious energy consumption funneling effect in EEUC and HEED-M. But UCOR has made great improvements by using rational unequally clustering method and routing mechanism: UCOR-2 and UCOR-3 successfully decreased the energy consumption in the top region of funnel by 20%. However, it's not the case that the more unequally the network is clustered, the effect is better. As shown in Fig.6, when we employ two unequal clusters, hot area of energy consumption is the region of cluster headers of the second cluster ring, and the energy consumption is far more than that of the top region in the funnel in EEUC and HEED-M. Besides, UCOR-3 generally eliminates the energy funneling effect.

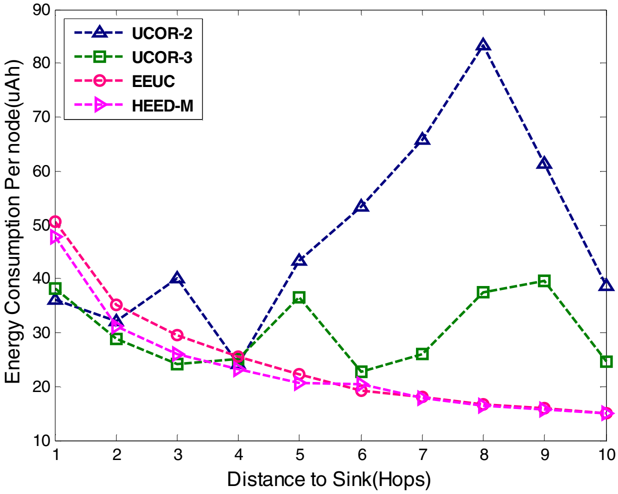


Fig. 6. Energy consumption in different regions in a single sampling period

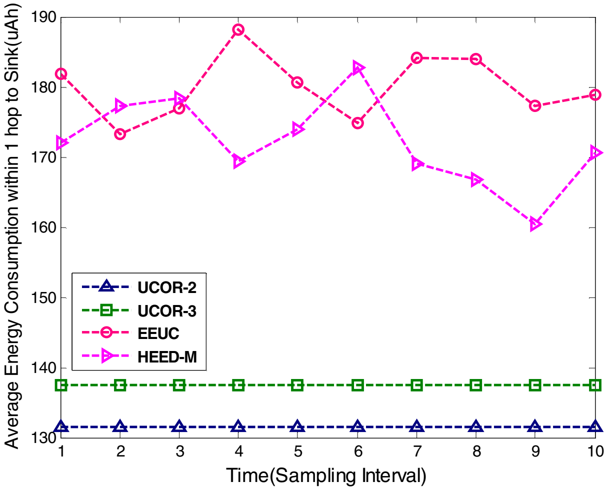


Fig. 7. Average energy consumption of nodes which are one-hop far from sink node

In addition, Fig.7 shows the average energy consumption variation of nodes in regions that are close to the sink node when time varies. In EEUC and HEED-M, since clustering process will repeat in each sampling period, network clustering topological structure changes severely, result in the routing varies frequently, further the conditions of data loss; retransmission and network congestion varies constantly. However, in UCOR, the static clustering method makes average energy consumption of nodes within the region generally constant.

5 Conclusions

This paper proposed UCOR protocol to solve energy consumption funneling effect problem in large-scale WSNs. In UCOR, network is unequally clustered based on characteristics of network traffic and energy consumption distribution, appropriate data aggregation methods are designed in different regions and opportunistic routing methods are adopted in various layers of network. By unequally clustering the network based on network load, energy consumption characteristics of sensor nodes in different region, UCOR protocol alleviated funneling effect. Simulation experiments demonstrated that UCOR protocol effectively lighten funneling effect and made energy consumption more balanced and lower, which prolongs network longevity and support for larger scale network.

Acknowledgements. This work was partially supported by National Key Basic Research Program of China “973 Project” (grant No. 2011CB707100), Program for Changjiang Scholars and Innovative Research Team in University (grant No. IRT1278), Basic Research Project of China Earthquake Administration (Grand No. 201101014), Director Foundation of the Institute of Seismology, China Earthquake Administration (Grand No. IS200726019, IS200956041).

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., et al.: Wireless sensor networks: a survey. Elsevier Computer Networks 38(4), 393–422 (2002)
2. Romer, K., Mattern, F.: The design space of wireless sensor networks. IEEE Wireless Communications 11(6), 54–61 (2004)
3. Ahn, G.S., Miluzzo, E.: Funneling-MAC: A Localized, Sink-oriented MAC for Boosting Fidelity in Sensor Networks. In: Proc. of the 4th ACM Conference on Embedded Networked Sensor Systems (SenSys 2006), New York, USA, pp. 293–306 (November 2006)
4. Chen, G., Li, C., Ye, M., Wu, J.: An Unequal Cluster-based Routing Protocol in Wireless Sensor Networks. Wireless Networks 15, 193–207 (2007)
5. Wu, X., Chen, G., Das, S.K.: On the Energy Hole Problem of Nonuniform Node Distribution in Wireless Sensor Networks. In: Proc. of 3rd IEEE International Conference on Mobile Ad hoc and Sensor Systems (MASS 2006), Vancouver, Canada, pp. 180–187 (October 2006)

6. Dutta, P., Feldmeier, M., Paradiso, J., Culler, D.: Energy Metering for Free: Augmenting Switching Regulators for Real-Time Monitoring. In: Proc. of International Conference on Information Processing in Sensor Networks (IPSN 2008), St. Louis, USA (2008)
7. Niu, X.G., Cui, L.: Throughput Capacity of Opportunistic Routing in Wireless Sensor Networks. International Journal of Distributed Sensor Networks 2010(148359), 1–9 (2010)
8. Younis, O., Fahmy, S.: HEED: A Hybrid, Energy-efficient, Distributed Clustering Approach for Ad-hoc Sensor Networks. IEEE Trans. on Mobile Computing 3(4), 660–669 (2004)
9. Li, C., Ye, M., Chen, G., Wu, J.: An Energy-efficient Unequal Clustering Mechanism for Wireless Sensor Networks. In: Proc. of International Conference on Mobile Adhoc and Sensor Systems (MASS 2005), pp. 1–8 (November 2005)
10. Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: An Application-specific Protocol Architecture for Wireless Miscrosensor Networks. IEEE Trans. on Wireless Communications 1(4), 660–670 (2002)
11. Soro, S., Heinzelman, W.B.: Prolonging the Lifetime of Wireless Sensor Networks via Unequal Clustering. In: Proc. of International Parallel and Distributed Processing Symposium (IPDPS 2005), pp. 1–8 (April 2005)
12. Siva, D., Daniel, C., Rolly, I., Abraham, O.: A centralized energy-efficient routing protocol for wireless sensor networks. IEEE Communications Magazine 43(3), 8–13 (2005)
13. Wang, Y., Yang, T., Zhang, D.: An energy efficient and balance hierarchical unequal clustering algorithm for large scale sensor network. Inform. Technol. J. 8(1), 28–38 (2009)
14. Nurhayati, N., Bayar, G., Lee, K.: An Energy Optimized Unequal Clustering Routing Protocol in Wireless Sensor Networks. In: Han, Y.-H., et al. (eds.) Ubiquitous Information Technologies and Applications. LNEE, vol. 214, pp. 611–619. Springer Science+Business Media, Dordrecht (2013)
15. Liu, A., Wu, X., Chen, Z., Gui, W.: Research on the energy hole problem based on unequal cluster-radius for wireless sensor networks. Elsevier Computer Communications 33, 302–321 (2010)

A Quadri-Stage Contention MAC Protocol with Opportunistic Network Coding Support for Underwater Acoustic Networks

Yun Liu and Bocheng Zhu

School of Electronics Engineering and Computer Science, Peking University, Beijing, China
lynuxly@gmail.com, zhubc@pku.edu.cn

Abstract. In this paper, a novel distributed TDMA medium access control (MAC) protocol for multi-hop underwater acoustic sensor networks (UASNs), termed opportunistic network coding supported quadri-stage contention protocol (NC-QSCP), has been proposed. The QSCP employs a concentrated contention procedure to form a transmission schedule, according to which nodes can perform collision free channel access in the following dedicated transmission stage. A contention probability calculation algorithm is designed to improve the efficiency of the data transmission, so that a heavy loaded node could acquire more channel resource. In the contention stage, the data flow information in 2-hop neighborhood is gathered, and can be exploited for both probability calculation and network coding opportunity discovery. When network coding is available, an XOR operation is applied to the packets with opposite directions. Both the contention probability calculation algorithm and the network coding scheme can remarkably improve the throughput and energy efficiency of QSCP protocol, especially in a tandem network carrying bidirectional traffics. The simulation illustrates that the opportunistic network coding achieves a 15% improvement in end-to-end throughput and reduces 20% energy consumption per delivered packet against QSCP.

Keywords: Underwater Acoustic Sensor Networks, TDMA, Medium Access Control, Opportunistic Network Coding.

1 Introduction

Underwater Acoustic Sensor Networks has been widely used in ocean resource exploration, disaster precaution, hydrologic monitoring and deep sea expedition. The excogitation of MAC protocols is rather difficult for UASNs, for the underwater acoustic channel has many poor transmission conditions, such as non-negligible propagation delay, noise, limited bandwidth, shadowing, and the problems of attenuation, absorption, multi-path and fading [1]. Besides, Energy efficiency is also an important factor in MAC protocol designing. TDMA is generally considered unsuitable for UASN because it is difficult to synchronize the timing of the nodes and the low throughput caused by long guard interval [2]. However, the advantages of TDMA protocol in energy efficiency are also remarkable. On one hand, the slotted transmission rarely

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 186–195, 2013.

collides, which saves the energy wasted on retransmission, on the other hand, since a sensor can easily know the transmission schedule, it can turn into sleep mode when it is idle, which saves the energy wasted on channel listening [3].

Many TDMA MAC protocols for UASNs have been presented recently. Slotted FAMA [4] applies time slotting to the original FAMA [5] protocol, which is based on carrier sense and handshake before data transmission. The slotted timing can improve the throughput and save energy. T-Lohi [3] divides the time into reservation period and data period. In the reservation period, a node uses an extra device to send a short tone to reserve the channel, so that in data period it can send data. T-Lohi achieves better throughput and energy performance than traditional CSMA, TDMA and ALOHA protocols. In TRAMA [6], nodes exchange priority information with its neighbors in a random access period, so that every node knows its priority among its 2-hop neighborhood, then the node with the highest priority shall transmit data in the following scheduled access period. PR-MAC [7] also exploits 2-hop neighborhood priority information to minimize the collision in reservation period, and prevent collision in data transmission scheme. PR-MAC has been tested in an underwater acoustic communication test-bed. These TDMA MAC protocols are all contention free, and perform well in throughput and energy efficiency. The features of the proposed MAC protocols, including slotted timing, separation of reservation and transmission, 2-hop neighborhood information exchange, auction system and energy saving strategy, enlightened the designation of original QSCP.

Research on network coding (NC) [8] has made great progress these years. The network coding can utilize the omnidirectional transmission property of wireless networks, save transmission times, improve the network capacity and reduce the energy consumption [9]. However, the interference, collision and topology change in wireless networks, especially in water acoustic sensor networks, can cause packet loss and decode failure. Thus the conflict free MAC protocols, particularly those with time-division mechanism, are suitable to perform network coding [10]. The time schedule of mac protocol can also be used to find network coding chances. Paper [11] builds a network coding based energy consumption model in underwater acoustic networks. Paper [12] provides a cross-layer designed opportunistic network coding scheme for UANs, which can reduce 24% of the energy consumption. Paper [13] proposes a concept of the combination of propagation delay based duplex and network coding in an ideal TDMA network, and the simulation shows that network coding can bring a significant benefit to throughput and energy efficiency of the UANs. In this paper we introduce an opportunistic network coding support to QSCP, making an attempt to exploit the MAC information to discover network coding opportunity. The simulation illustrates that the opportunistic network coding support can remarkably improve the throughput and energy performance.

The rest of this paper is organized as follows: Section 2 introduces the frame structure, contention procedure, contention probability calculation algorithm and NC supporting of NC-QSCP; Section 3 evaluates the performance of NC-QSCP in a tandem network simulation; Section 4 draws a conclusion of this paper and discusses the future directions of this research.

2 Protocol Description

2.1 Frame Structure

In the QSCP process, time is divided into several time frames, including contention frames and information frames. In the contention frame (CF), nodes contend to generate a transmission schedule, which is used in the following L information frames (IF). One CF and L IFs constitute a QSCP period. Both CF and IF include N time slots, called contention slot (CS) and information slot (IS) respectively, and the winner in the n^{th} contention slot has the privilege to transmit data in the n^{th} information slot. A node contends with its 2-hop neighbors for M contention cycles (CC) in a CS, and a broadcasting cycle (BC) is employed to broadcast the contention result at the end of the CS. In each CC, nodes invoke a four stage contention procedure to attempt to win the transmission chance, and the protocol's name (quadri-stage contention protocol) comes from this procedure.

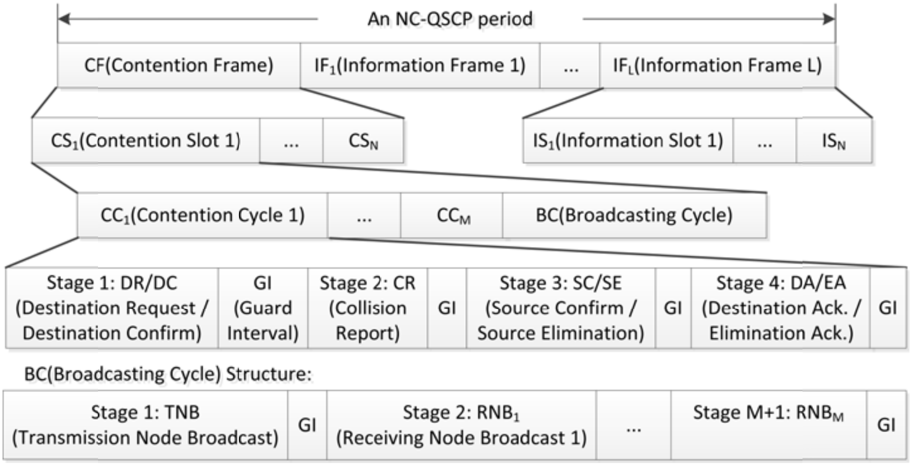

An NC-QSCP period
CF(Contention Frame)
IF₁(Information Frame 1)
...
IF_L(Information Frame L)
CS₁(Contention Slot 1)
...
CS_N
IS₁(Information Slot 1)
...
IS_N
CC₁(Contention Cycle 1)
...
CC_M
BC(Broadcasting Cycle)
Stage 1: DR/DC (Destination Request / Destination Confirm)
GI (Guard Interval)
Stage 2: CR (Collision Report)
GI
Stage 3: SC/SE (Source Confirm / Source Elimination)
GI
Stage 4: DA/EA (Destination Ack. / Elimination Ack.)
GI
BC(Broadcasting Cycle) Structure:
Stage 1: TNB (Transmission Node Broadcast)
GI
Stage 2: RNB₁ (Receiving Node Broadcast 1)
...
Stage M+1: RNB_M
GI

Fig. 1. NC-QSCP frame structure

In a stage of a contention cycle, some certain nodes shall send control packets, at the same time other node shall listen. A control packet contains some information, including control packet type, two certain MAC addresses, the local transmission queue length, a time stamp and a CRC. The receiver can use the two MAC addresses in the packet to deduct the data flow information of all its 2-hop neighbors. The transmission queue length is for contention probability calculation. The time stamp is used for synchronizing. QSCP is a TDMA protocol, thus time synchronizing is required among the nodes. However, since a node only contends within its 2-hop neighborhood, a global time synchronizing is not necessary. When a node collects the clock information of its neighbors, its local clock can be adjusted to the average time of its neighbors.

2.2 Quadri-Stage Contention Procedure

According to the contention status during the contention cycle, nodes are labeled as idle node (IDN), pre-requesting node (PRN), source node (SRN), destination node (DTN), affected node (AN), transmitting node (TN), transmitting nodes with NC packet (TN/NC), receiving node (RN) and blocked node (BN). When a CS begins, all nodes are reset to IDN state. If the MAC layer transmitting queue holds some data packets, or an NC opportunity has been detected, a node shall changes its status to PRN. Also, at the beginning of the CS, PRNs have to re-calculate the contention probability using the neighbors' queue length information which is collected in the previous CS. The PRNs use the contention probability in the beginning of DR/DC stage to decide whether to send a DR packet or to restrain itself and listen to the channel. The state transition diagram is shown in fig. 2. The 4 stages of a CC are described as follows:

In the first stage, the destination request / destination confirmation (DR/DC) stage, a node shall broadcast a DR or DC frame at the probability of P. DR is broadcasted by a PRN, and DC broadcasted by a TN. If a node has some NC chance, the DR must carry an NC label. Both DR and DC frames contain the sender's address and the destination node address (DtA) of the pending data packet. If a listener receives multiple frames, a collision can be detected, and must be reported in the next stage. The DC frame intends to confirm the reservation result. When a TN receives a DC frame and the DtA of that DC frame matches the TN's address, then the TN is blocked and shall change its status to BN. The algorithm of probability P's calculation will be discussed later. As the NC packet is transmitted preferentially, if a PRN receives a DR(NC) packet, it shall turn into BN directly.

In the second stage, the collision report (CR) stage, node which has detected a collision in the first stage shall broadcast a CR frame. The CR frame contains the address of the reporting node, called collision site address. If a PRN/TN receives one or more CR frames, it will know that the destination node has been affected by another node, and has to stop doing anything in the following stages of the current CC. In addition, if the current cycle is not the first CC of the CS, when RN, AN, or TN receives a DR/DC, it must also broadcast a CR frame.

In the third stage, source confirmation / source elimination (SC/SE) stage, if no CR frame has been received by a PRN, it shall broadcast an SC frame, which contains the sender's address and the final destination address (FdA) of the pending data packet. The RN can use FdA to detect and NC chance. At the same time, if a BN does not receive any frame in stage 1 and stage 2, it shall broadcast an SE frame to inform other nodes that it shall withdraw the previous reservation, so that its RN can accept another reservation later.

The fourth stage is the destination acknowledgement / elimination acknowledgment (DA/EA) stage. If a node is available to become an RN, it shall reply a DA frame to the TN. In another hand, if a PRN had not broadcasted a DR but received a DA frame from its destination node, it shall become a BN. A DA frame also carries the yonder address of this reservation. The yonder address indicates the node after next hop of that data packet, and it could be found in the RN's route table. Otherwise, if a node has broadcasted DR and SC but receives no DA, it shall turn back to be a PRN instead of TN. Such situation may be caused by either of the conditions: (1) the

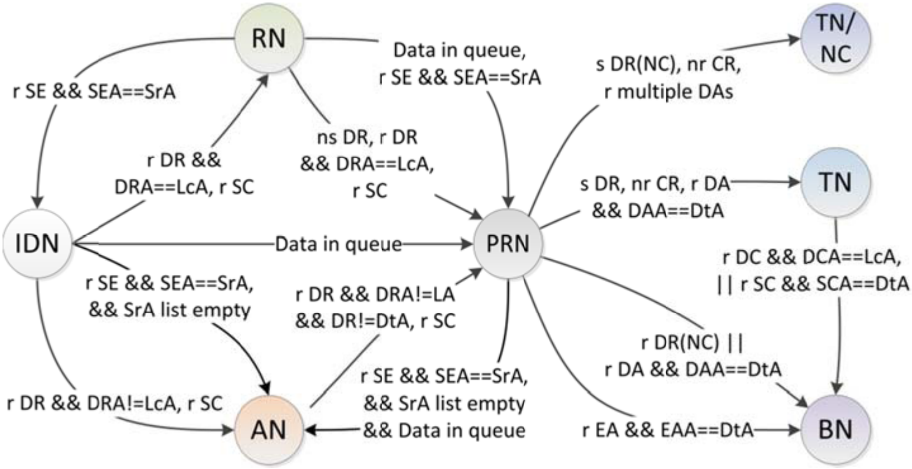

RN
TN/ NC
TN
IDN
PRN
AN
BN
r SE && SEA==SrA
Data in queue, r SE && SEA==SrA
s DR(NC), nr CR, r multiple DAs
r DR && DRA==LcA, r SC
ns DR, r DR && DRA==LcA, r SC
s DR, nr CR, r DA && DAA==DtA
Data in queue
r SE && SEA==SrA, && SrA list empty
r DR && DRA!=LA && DR!=DtA, r SC
r DC && DCA==LcA, || r SC && SCA==DtA
r DR(NC) || r DA && DAA==DtA
r SE && SEA==SrA, && SrA list empty && Data in queue
r DR && DRA!=LcA, r SC
r EA && EAA==DtA

Fig. 2. State transition diagram during a contention cycle

node is an isolated one, the destination node is not a neighbor; (2) the destination node has already become an RN of another TN. In the other hand, if an RN has just received an SE frame, it shall reply an EA frame in this stage. If a BN did not send SE in the third stage but received an EA from its destination node, it can recover itself to be a PRN.

Fig. 2 illustrates the state transition and control message exchange procedure. In the arrow's label, letter "r" represents "received", "s" stands for "sent", "nr" and "ns" means "not received" and "not sent" respectively. Capital letter "A" means "address", for instance, "DRA" means the address field carried by the DR frame. DtA, LcA and SrA are short for destination address, local address and source address respectively. DtA is used by nodes that have some data to transmit, and it points to the next hop of this packet according to the route table. With regard to an RN, SrA is the address of its corresponding TN. SrA list is an address list of all the affecting nodes for an AN.

In the BC, which is the conclusion cycle of CS, the TNs shall check their reservation status, including the index of reserved IS, the transmitting queue length, the destination address and yonder address, and then broadcast the information to their neighbors in the TNB stage. In the following M RNB stages, the RNs shall broadcast an RNB packet successively in the next M RNB stages. If a node becomes an RN in the mth cycle, it shall broadcast at the (m+1)th stage of BC. After the TNB, most nodes shall know the topology and queue length of the 2-hop neighbors, and new probabilities are calculated and used for next contention slot. At the same moment, the NC chances are also detected by checking the source and yonder addresses of previous reservations and the intermediate node shall mark the RNB packet to inform its neighbors that it will take part in the contention of next CS, to reserve the time slot for transmitting an NC packet.

2.3 NC Supporting

NC-QSCP supports opportunistic network coding within 2-hop neighborhood. The function requires route information support of the network layer. As shown in fig. 3, when node A and B exchange packet a1 and b1 via intermediate node M, an NC chance is detected, and M shall reserve a slot for transmitting the NC packet (a1+b1) to node A and node B. The NC chance is detected by analyzing the TNB's address field yonder address, which is acquired in the four stage handshake progress.

Suppose node A holds a data packet a1 whose final destination is node D, it already knows that next hop is M according to the route table. Then after a successful four stage handshake with node M, it shall know that packet a1's next 2 hops are M and B. Then in the BC stage, node M confirms that it shall forward packet a1 to B. In another CS, node M also finds that it shall forward packet b1 to A. Then node M shall combine the two forward operations to one NC broadcast. Node M shall reserve a slot for broadcasting packet a1+b1 in the next CS, and node A and node B shall prepare a decode operation when the NC packet comes.

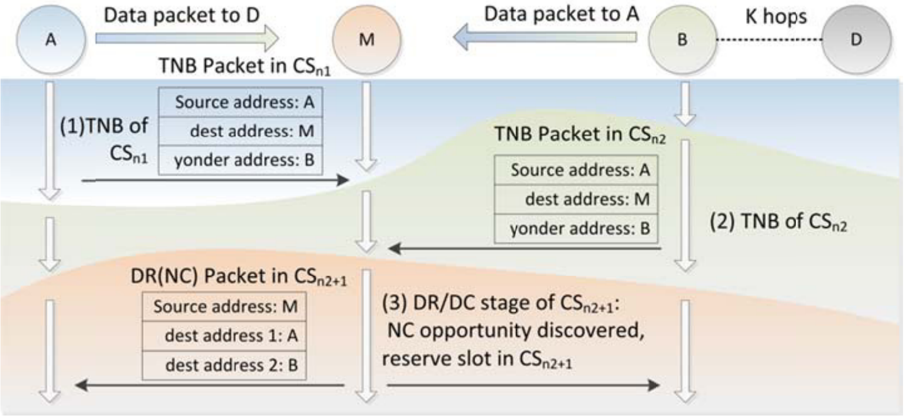


Fig. 3. NC opportunity discovery during QSCP handshake procedure

2.4 Contention Probability Calculation

In the DR/DC stage, a PRN shall broadcast DR packet at the probability of *P*. The algorithm of generating *P* is discussed in this section. It is noticed that a node with longer data queue requires more time slots. Then the successful reservation probability shall be proportional to the data queue length, so that the time slot reservation result is fair for all nodes. We use *Lq* to represent the data queue length, and *Psuc* for successful reservation probability, then in a 2-hop neighborhood, we have:

$$\frac{Psuc_0}{Lq_0} = \frac{Psuc_1}{Lq_1} = \cdots = \frac{Psuc_K}{Lq_K} \tag{1}$$

Where *K* is the number of nodes in the 2-hop neighborhood.

A node makes a successful reservation when it broadcasts DR packet at the probability of P_i, and at the same time other nodes fail to do so. Then:

$$Psuc_i = \frac{P_i}{(1-P_i)} \prod_{k=1}^{K} (1 - P_k)\,, [1 \le k \le K] \tag{2}$$

If any node among the 2-hop neighborhood has made a successful reservation, the contention procedure of the current CS is finished, so the probability of any node's successful reservation shall be the sum of all *Psuc* of the 2-hop neighborhood:

$$Pa := \sum_{k=1}^{K} Psuc_k \tag{3}$$

We shall find a group of P_i to maximize *Pa*, so that the contention can be effective. Let C_i be the normalized Lq_i:

$$C_i := \frac{Rn_i}{\sum_{j=1}^{K} Rn_j}, [1 \le i \le K] \tag{4}$$

Then from (1), (2), (3) and (4) it could be derived that:

$$Pa \le \prod_{j=1}^{K} (1 - C_j)^{1-C_j} \tag{5}$$

When:

$$P_i = 1 - \frac{1}{\left(1 + C_i \prod_{j=1}^{K} (1 - C_j)^{-C_j}\right)},\ [1 \le i \le K] \tag{6}$$

Pa can reach its max value.

So after a node gathered the *Lq* information of its 2-hop neighbors, it can apply equation (4) and (6) to calculate its contention probability, so that when it takes part in the contention, its success probability is proportional to its data queue length.

3 Simulation Results

The NC-QSCP is evaluated by OMNeT++ [14] simulations. A tandem network with 10 nodes (shown in figure 1) is employed to test the protocol, which is the same as the bidirectional traffic flowing linearly-arranged underwater network in [12].

Table 1. Common Parameters

Transmission Range	300m
Routing Protocol	AODV
MAC Queue Length	30 pkt.
Simulation Duration	900 min
Phy-Layer Bandwidth	1Kbps
Carrier Frequency	50KHz
Power Spent in TX mode	0.6mW
Power Spent in RX mode	0.3mW
Power Spent when idle	0mW

Table 2. NC-QSCP Parameters

Section	Structure	Time
Period	1 CF + 1 IF	260s
CF/IF	10 CS/IS	130s
IS	1500 bytes	13s
CS	5 CC + 1 BC	13s
CC	4 Stages	2s
BC	6 Stages	3s
Stage	128 bits + GI	500ms
Guard Int.	372ms	372ms

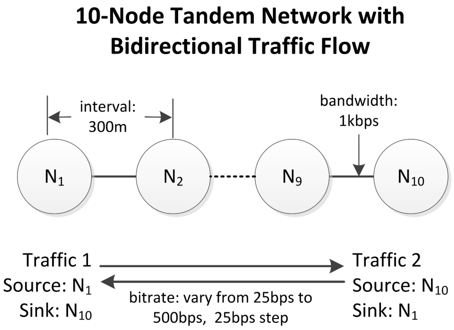


Fig. 4. Tandem Network for Simulation

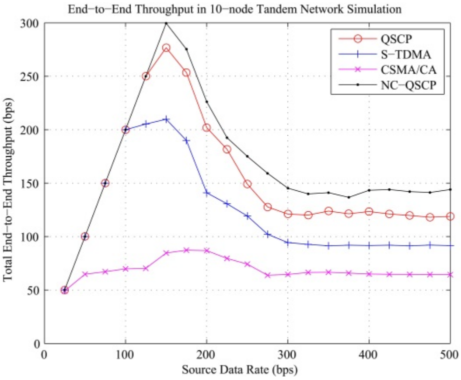


Fig. 5. Throughput Performance Comparison

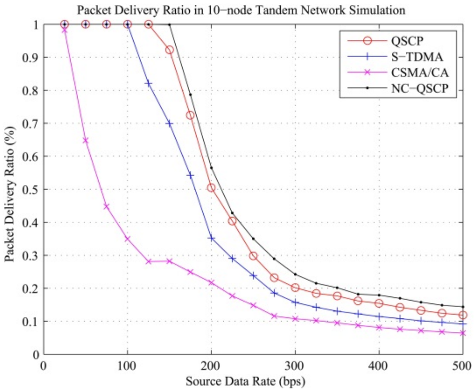


Fig. 6. Packet Delivery Ratio Comparison

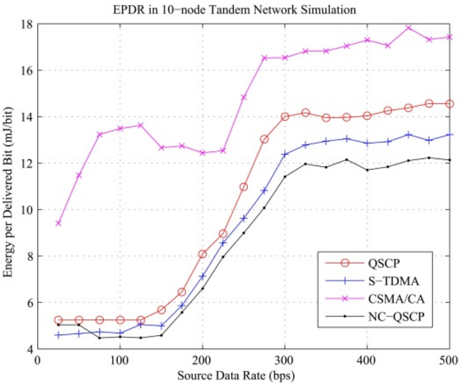


Fig. 7. Energy Consumption Performance

As shown in fig. 4, N1 and N10 are transmitting data packet to each other via 8 intermediate nodes. The nodes communicate with each other by broadcasting acoustic waves in the water, and the propagation speed is 1500m/s. Table 1 illustrates the common parameters of all the simulated protocols, including QSCP, NC-QSCP, CSMA/CA and S-TDMA. We assume that the channel is noiseless, so that the packet lost is only caused by collisions.

We focus on the performance of end-to-end throughput, packet delivery ratio and energy efficiency. Fig. 5 illustrates the throughput performance of the 4 protocols. In the acoustic sensor networks, due to the non-negligible propagation delay and low transmission rate, carrier sense become unreliable, collisions cannot be avoided, thus CSMA protocols does not perform well. The contention probability calculation algorithm of QSCP can provide about 25% improvement of throughput performance to ordinary TDMA protocol, and the opportunistic NC operation can provide 10% to

20% more than QSCP. It is noticed that the throughput declines when source rate is larger than 150bps. The reason is that when source data rate rises, the bandwidth become insufficient, then some data packets has to be dropped. At this time, a long traveled data packet ought to be kept and a new generated packet shall be dropped, because dropping a packet means wasting the resources spent on it, and a long travelled packet has yet consumed more resource than other packets. However, MAC does not know the hop information of a packet, so it can only randomly drop packets. So when the two data flow rate is larger than 150bps, the total bitrate is about 300bps, which is the upper bound of tandem network throughput [15], and the throughput begins to descend while more and more long traveled packets are randomly dropped.

We use the "energy per delivered bit (EPDB)" index to evaluate the energy efficiency of protocols. EPDB is the quotient of total energy consumption of the network (in joule) and the number of bits successfully reached the final destination's application layer. EPDB includes not only the energy used in control overhead and data transmission, but also the energy wasted in collision and retransmission. A protocol with lower EPDB would be more efficient in energy saving. Fig. 7 gives the comparison of EPDB among the 4 protocols. It is noticed that TDMA protocols is much more efficiency in energy than CSMA protocols, and network coding can improve energy performance. However, it is also discovered that when source rate is low, QSCP and NC-QSCP spends more energy in contention frames, which caused the disadvantage against S-TDMA.

4 Conclusion

A novel opportunistic network coding supported probability driven quadri-stage contention protocol (NC-QSCP) for underwater acoustic sensor networks has been proposed. The protocol employs a concentrated contention phase to arrange time slot for data transmission, opportunistic network coding is incorporated in the contention phase, exploiting the 2-hop neighborhood data flow information to discover the opportunity. A contention probability calculation algorithm is also provided to improve the end-to-end throughput, packet delivery ratio and energy efficiency. The protocol is suitable for a linear topology sensor network carrying bidirectional traffics. Simulation results prove that the contention probability calculation algorithm and the opportunistic network coding scheme can remarkably improve the throughput and energy performances, especially when network load is heavy.

In the future, we plan to introduce a generic linear network into QSCP protocol, make it suitable for arbitrary topology sensor networks.

References

1. Makhija, D., Kumaraswamy, P., Roy, R.: Challenges and Design of Mac Protocol for Underwater Acoustic Sensor Networks. In: 4th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, pp. 1–6 (2006)

2. Sozer, E.M., Stojanovic, M., Proakis, J.G.: Underwater acoustic networks. IEEE Journal of Oceanic Engineering 25(1), 72–83 (2000)
3. Syed, A.A., Ye, W., Heidemann, J.: Comparison and Evaluation of the T-Lohi MAC for Underwater Acoustic Sensor Networks. IEEE Journal on Selected Areas in Communications 26(9), 1731–1743 (2008)
4. Molins, M., Stojanovic, M.: Slotted FAMA: a MAC protocol for underwater acoustic networks. In: OCEANS 2006 Asia–Pacific, pp. 1–7 (2006)
5. Garcia-Luna-Aceves, J.J., Fullmer, C.L.: Performance of floor acquisition multiple access in ad-hoc networks. In: 3rd IEEE Symposium on Computers and Communications, pp. 63–68 (1998)
6. Chih-Lin, I., Pollini, G.P.: The tree-search resource auction multiple access (TRAMA) protocol for wireless personal communications. In: IEEE 44th Vehichular Technology Conference, vol. 2, pp. 1170–1174 (1994)
7. Hui-Jin, C., Jung-Il, N., Nam-Yeol, Y., et al.: Contention free MAC protocol based on priority in underwater acoustic communication. In: OCEANS 2011, Spain, pp. 1–7 (2011)
8. Ahlswede, R., Ning, C.: Network information flow. IEEE Transactions on Information Theory 46(4), 1204–1216 (2000)
9. Chirdchoo, N., Chitre, M., Wee-Seng, S.: A study on network coding in underwater networks. In: OCEANS 2010, pp. 1–8 (2010)
10. Otsuki, N., Sugiyama, T.: Performance Evaluation of TDMA Based Wireless Network Coding Prototype System. In: IEEE 2012 Vehicular Technology Conference (VTC Fall), pp. 1–5 (2012)
11. Lucani, D.E., Medard, M., Stojanovic, M.: Underwater Acoustic Networks: Channel Models and Network Coding Based Lower Bound to Transmission Power for Multicast. IEEE Journal on Selected Areas in Communications 26(9), 1708–1719 (2008)
12. Haojie, Z., Valera, A., Zhi, A.E., Lee, P.W.Q., Tan, H.P.: Opportunistic XOR network coding for multihop data delivery in underwater acoustic networks. In: OCEANS 2011, Spain, pp. 1–7 (2011)
13. Palacios, R., Heide, J., Fitzek, F.H.P., Granelli, F.: Design and performance evaluation of underwater data dissemination strategies using Interference Avoidance and Network Coding. In: IEEE International Conference on Communications (ICC), pp. 1410–1415 (2012)
14. `http://www.omnetpp.org/`
15. Xiao, Y., Peng, M., Gibson, J., et al.: Tight Performance Bounds of Multihop Fair Access for MAC Protocols in Wireless Sensor Networks and Underwater Sensor Networks. IEEE Transaction on Mobile Computing 11(10), 1538–1554 (2012)

Fast Encryption of JPEG 2000 Images in Wireless Multimedia Sensor Networks

Tao Xiang[1], Chenyun Yu[1], and Fei Chen[2]

[1] College of Computer Science, Chongqing University, Chongqing 400044, China
[2] Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Hong Kong, China
txiang@cqu.edu.cn,
yuchenyun0320@163.com,
fchen@cse.cuhk.edu.hk

Abstract. A selective encryption algorithm joint with compression coding is proposed to protect JPEG 2000 images in wireless multimedia sensor networks (WMSN). The algorithm selectively encrypts the lookup table of probability model in MQ coding. As the size of lookup table is fixed and only such one table is used for an image, the proposed algorithm is fairly efficient and thus can perform fast encryption on large volume of JPEG 2000 images in WMSN. Experimental results and their analysis show that the algorithm is secure and energy saving, meanwhile, it does not impair the compression performance of JPEG 2000 coding obviously.

1 Introduction

With the fast development and the widespread application of wireless sensor networks (WSN), more and more data are collected by sensor nodes [1]. They not only include scalar data such as temperature and humidity, but also contain vector information such as images and videos. Wireless multimedia sensor networks (WMSN) refer to the WSN that mainly deal with multimedia data [2], and therefore they inherit many characteristics such as resource limitation from WSN. WMSN have been proposed and drawn the immediate attention of the research community with the availability of low-cost small-scale imaging sensors, CMOS cameras, microphones, which may ubiquitously capture multimedia content from the field. WMSN not only enhance existing sensor network applications such as tracking, home automation, and environmental monitoring, but they also enable several new applications such as surveillance, traffic congestion monitoring, health care, and industrial process control.

The security of WMSN is a primary issue since the camera nodes are usually deployed in public environment and the data are transmitted in wireless, however, traditional ciphers provide little help on protecting multimedia data in WMSN for multifold reasons. First, massive volumes of persistently generated multimedia data make it infeasible to adopt traditional computationally intensive ciphers. Second, the hardware limitation of sensor node, as well as its

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 196–205, 2013.

energy sensitivity, requires the overhead reduction on encryption. Last, to meet QoS requirement on multimedia delivery, encryption process should be designed as fast as possible to alleviate its impact on realtime transmission.

Generally speaking, there are two promising ways to alleviate the conflict between security requirement and image processing and transmission. One is selective encryption [3–9], and its principle is reducing the amount of data to be encrypted by only selectively encrypting a portion of data which is defined as important. The other is joint compression and encryption [10–14, 8, 7, 15–21, 9], and it saves the overhead by combining the process of data encryption and data compression into single one step. There are some related work along these two lines for protecting multimedia data in WMSN [22–26], and only a portion of data is encrypted to save the energy for sensor nodes. That is to say, the encryption overhead is proportional to the total of data volume, and more computational cost is needed on encryption when dealing with massive data.

In this paper, we combine the idea of selective encryption with joint compression and encryption, and propose a fast selective encryption algorithm for secure JPEG 2000 coding. Different from most existing selective encryption schemes, the volume of selectively encrypted data is constant for an image no matter what the size of it is. As only a tiny part of fixed-length data is encrypted regardless of the size of whole image file, the algorithm is extremely fast and is suitable for protecting JPEG 2000 images in WMSN. Experimental results and their analysis show that the proposed algorithm achieves a good tradeoff between security and efficiency.

The rest of this paper is organized as follows. Section 2 gives a brief introduction about JPEG 2000 and MQ coder. In Section 4, experimental results on the proposed algorithm and its analysis are provided. Finally, Section 5 concludes the paper.

2 JPEG 2000 and MQ Coder

JPEG 2000 is a new image compression standard, and it acts as an update of the wide-spread JPEG image standard [9, 27]. Like JPEG coding, the JPEG 2000 coding process includes four phases: pre-processing, orthogonal transform, quantization, and entropy coding; but it uses discrete wavelet transform (DWT) in orthogonal transform and MQ coder as the entropy coder. JPEG 2000 offers numerous advantages over JPEG standard, and one prominent advantage is that it offers higher compression ratio, especially in low bit-rate. It is reported that JPEG 2000 outperforms JPEG by more than 30% in compression at 0.5bpp or less [9], which makes it a good candidate in resource constrained environments such as WMSN.

JPEG 2000 performs entropy coding on bit plane by MQ coder, and generates highly compressed data bit. MQ coder is essentially context-based adaptive binary arithmetic coding (AC), and it is multiplication-free in order to accelerate the coding. As shown in Fig. 1, upon receiving a bit to be encoded (D), MQ coder first decides whether it is the least probable symbol (LPS) or the most probable symbol (MPS) based on its probability. The probability estimation of LPS (Q_e) and its update are provided by the standard as a lookup table with four fields as shown in Fig. 2. Then it updates the current coding interval (A) accordingly.

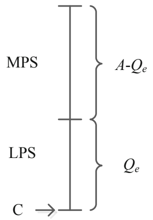
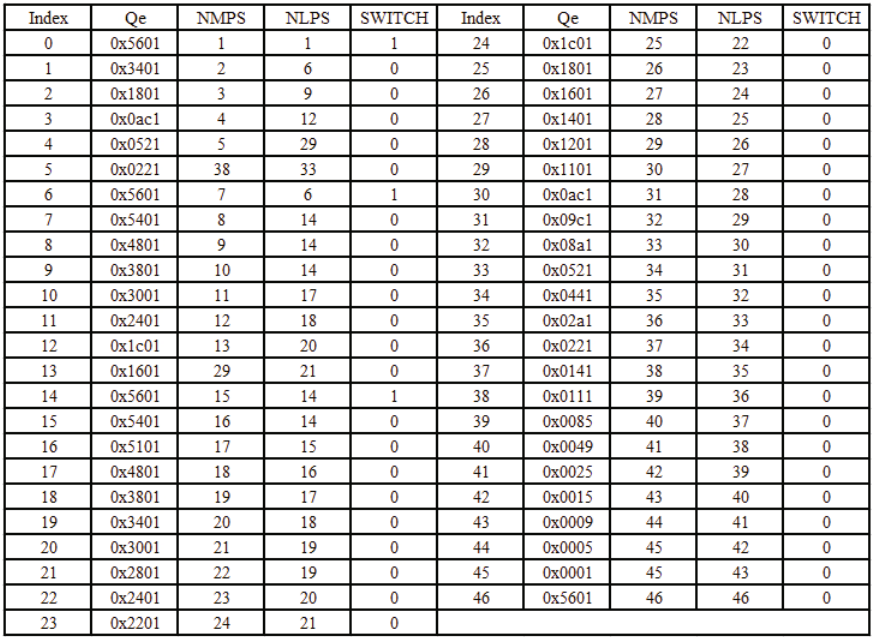


Fig. 1. MQ coder



Index	Qe	NMPS	NLPS	SWITCH	Index	Qe	NMPS	NLPS	SWITCH
0	0x5601	1	1	1	24	0x1c01	25	22	0
1	0x3401	2	6	0	25	0x1801	26	23	0
2	0x1801	3	9	0	26	0x1601	27	24	0
3	0x0ac1	4	12	0	27	0x1401	28	25	0
4	0x0521	5	29	0	28	0x1201	29	26	0
5	0x0221	38	33	0	29	0x1101	30	27	0
6	0x5601	7	6	1	30	0x0ac1	31	28	0
7	0x5401	8	14	0	31	0x09c1	32	29	0
8	0x4801	9	14	0	32	0x08a1	33	30	0
9	0x3801	10	14	0	33	0x0521	34	31	0
10	0x3001	11	17	0	34	0x0441	35	32	0
11	0x2401	12	18	0	35	0x02a1	36	33	0
12	0x1c01	13	20	0	36	0x0221	37	34	0
13	0x1601	29	21	0	37	0x0141	38	35	0
14	0x5601	15	14	1	38	0x0111	39	36	0
15	0x5401	16	14	0	39	0x0085	40	37	0
16	0x5101	17	15	0	40	0x0049	41	38	0
17	0x4801	18	16	0	41	0x0025	42	39	0
18	0x3801	19	17	0	42	0x0015	43	40	0
19	0x3401	20	18	0	43	0x0009	44	41	0
20	0x3001	21	19	0	44	0x0005	45	42	0
21	0x2801	22	19	0	45	0x0001	45	43	0
22	0x2401	23	20	0	46	0x5601	46	46	0
23	0x2201	24	21	0					

Fig. 2. The lookup table of Q_e estimation defined in JPEG 2000

3 The Proposed Fast Encryption Algorithm

Based on the above statements, we can find that the adaptivity of MQ coder is decided by the value of Q_e in lookup table as shown in Fig. 2. For example, if LPS is coded and the interval is renormalized, the value of Q_e should be increased;

if MPS is coded and the interval is renormalized, Q_e should be decreased. The updating of Q_e is determined by the updating of entry index in the lookup table. In this manner, the probability model can accurately reflects the statistical probabilities of the input, and the coding result will approach the minimum entropy. The MQ decoder imitates this process to recover the source message. That is to say, the values of Q_e in coding and decoding are identical. If the value of Q_e has tiny difference in decoding process, the current coding interval will change. What's more, the iteration of coding procedures will magnify this difference and finally make the decoded output totally different.

Based on this fact, the basic idea of our proposed encryption algorithm is altering the values of Q_e in lookup table as shown in Fig. 2. Specifically, a secret disturbance is applied on Q_e for each index, and the disturbance is tuned within an appropriate range to get a good tradeoff between security strength and compression performance. The encryption operation can be formulated as below:

$$Q_e = Q_e + r \tag{1}$$

where r is a random number uniformly distributed in $[0, R]$, and it is generated by a cryptographically secure pseudorandom number generator (PRNG) with a secret key; R is the threshold controlling the impact on compression performance. In this manner, MQ coder uses a secure adaptive statistical model for JPEG 2000 encoding. MQ coder cannot replay this disturbance without the correct secret key, and thus make the JPEG 2000 decoded image far different from its plain image.

The proposed algorithm is a selective encryption joint with JPEG 2000 coding. Its main advantage over other existing selective image encryption algorithms is that the encryption overhead is not proportional to the size of plain images, i.e. the volume of selected data to be encrypted is fixed regardless of the size of plain images, as the size of lookup table for Q_e is fixed and there is only such one table per image. What's more, the lookup table contains 47 entries, which means only 47 random numbers should be generated during the encryption process. For these reasons, the proposed encryption scheme is efficient, especially when dealing with massive image data in real time, and can serve as a fast lightweight encryption algorithm in WMSN.

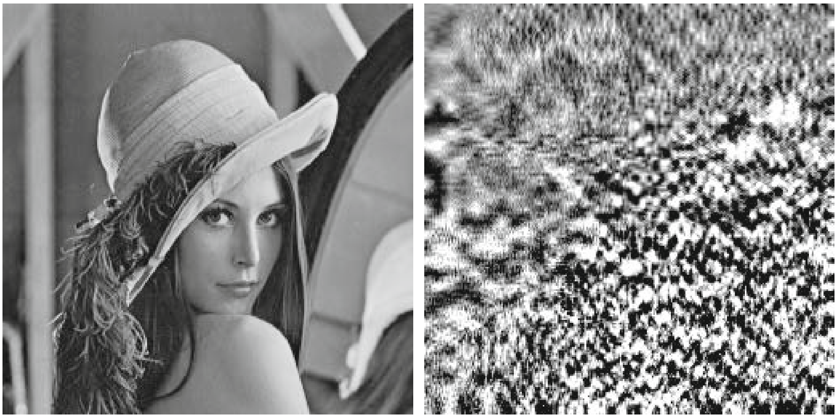
4 Experimental Results and Analysis

This section describes experimental results and their analysis in order to validate and evaluate the proposed algorithm. A 256×256 gray-level Lena image is taken as the plain image. The compression ratio of JPEG 2000 is set to 0.1 for a balance between image quality and image size. RC4 is adopted as the PRNG to generate random number r. The threshold R is set to 0x0600 to get a tradeoff between security strength and compression performance.

4.1 Encryption Results

We need to validate the encryption effect of the proposed algorithm since only a fixed tiny portion of data is selected to be encrypted. Fig. 3 demonstrates the visual difference between plain image and its encrypted image. The plain image is shown in Fig. 3(a), it is encrypted by the proposed algorithm and the encrypted image is given in Fig. 3(b). It is clear that the content of encrypted image is confused, and nothing intelligible about the plain image can be inferred from it.

To measure the quality of encrypted image and its difference from the plain image, we calculate the peak signal-to-noise ratio (PSNR) of these two images, and the results are given in Table 1. From Table 1, the PSNR of encrypted image is only 8.57, indicating that its quality is fairly low. Compared with the PSNR of plain image 34.89, it is easy to understand that the content of plain image is well confused in the encrypted image. Based on these results, the proposed algorithm is proved to be capable of protecting the entire image by only selectively encrypting the values of Q_e in the lookup table.



(a) Lena (b) Encrypted Lena

Fig. 3. Lena and its encrypted image

4.2 Security Analysis

Resistance to Brute-Force Attack: To avoid adversaries from guessing the key using brute-force attack, the key space of a cipher should be designed to be sufficiently large. In our proposed algorithm, the key space is well guaranteed by the key space of cryptographically secure PRNG. For example, in the experiments, RC4 is utilized as the PRNG, and thus the maximum key space is 2^{256}.

Table 1. Encryption results on Lena

Image	PSNR
JPEG 2000 coded image	34.89
JPEG 2000 coded image with encryption	8.57

Even so, there is another worry about the possibility of brute-force attack on the proposed algorithm since the lookup table of Q_e only contains 47 entries, which means only 47 values are encrypted. The attacker could directly recover the content of this table if the magnitude of disturbance on Q_e, i.e. R, is small. Therefore, R should be set as large as possible, as long as the compression performance is not significantly degraded. In our experiments, R is set to 0x0600, so the space of ciphertext is $\texttt{0x0600}^{47} \approx 2^{498}$. This is a fairly large to suppress the feasibility of brute-force attack. At the same time, the compression performance is not obviously impaired (please refer to subsection 4.3 for the detail).

Based on the above analysis, we can see that the proposed encryption algorithm can resistant to brute-force attack by proper selection of a cryptographically secure PRNG and the value of R.

Key Sensitivity: Key sensitivity is required for a good encryption algorithm since the security of a cryptosystem only depends on the secret of the key. Because MQ is adaptive AC in essence, with coding of incoming symbols, the statistical model can be adaptively adjusted to reflect the real distribution of the source. In other words, MQ is not sensitive to the initial condition of probability model determined by the value of Q_e. In our proposed algorithm, the key is used to generate a keystream by PRNG, and then the keystream is utilized to encrypt/decrypt Q_e. Therefore, we need to exam the sensitivity of the key.

In our proposed algorithm, all the values of Q_e in the lookup table are disturbed. Once a binary symbol is encoded and the interval is renormalized, MQ coder will locate a new entry in the lookup table by the index number to update the value of Q_e, so the encryption algorithm does not only change the initial probability model, but also disturbs it on the fly. In this manner, the encryption algorithm is sensitive to the key. In the experiments, we randomly generate 3000 keys that are uniformly distributed in the key space, and take only one of them as the correct key to encrypt the plain image. The PSNRs of decrypted images by all these keys are plotted in Fig. 4, and the PSNRs obtained by incorrect keys are all below 10. It shows that little change in secret key seriously affects the quality of the decrypted image and makes it unintelligible. Therefore, the key sensitivity of the proposed encryption algorithm is guaranteed.

4.3 Compression Performance

As the proposed encryption algorithm is combined with JPEG 2000 coding process, it should not have significant negative effect on the compression performance of JPEG 2000 coding. In the proposed algorithm, the encryption is

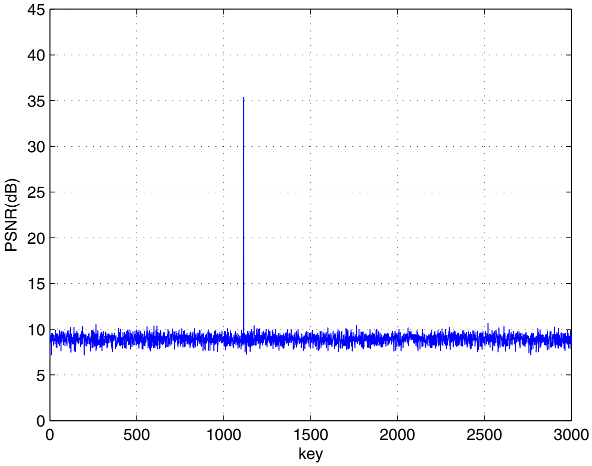


Fig. 4. PSNR obtained by decoding all possible combinations of the keys

performed by securely changing the value of Q_e, and the probability model may different from the standard MQ coding. This alteration may make the probability model deviate from or approximate to the real statistics of input source, so the compression performance of the encryption algorithm should be investigated.

We tabulate the length of compressed code stream by standard JPEG 2000 and the proposed encryption algorithm in Table 2, and find that the encryption does not impair the compression performance of JPEG 2000 obviously. Actually, this result is attributed to the appropriate selection of threshold R since it controls the magnitude of disturbance on Q_e. The greater value Q_e takes, the more disturbance it will make on the probability model of MQ and thus has higher probability to impair the compression performance of MQ. In our experiments, it is found that if R takes a value greater than 0x1000, the compression ratio after encryption will be obviously degraded. For this reason, we set R as 0x0600 to get a good balance between security and compression performance. In practice, other value could also be configured to support different application scenario.

Table 2. Compression performance comparison on Lena

Image	Code stream length (byte)
JPEG 2000 coded image	25905
JPEG 2000 coded image with encryption	25852

4.4 Energy Consumption in WMSN

Because the algorithm is proposed to protect JPEG 2000 images in WMSN and sensor nodes in WMSN are energy sensitive, it is necessary to analyze the energy consumption of the proposed encryption algorithm. Since the size of the lookup table is fixed, the encryption efficiency of the proposed algorithm is better than many existing schemes where the data to be encrypted are proportional to the image size. We implement the proposed algorithm in OMNeT++ [28]. The sensor node is configured with PXA255 as its processor (400MHz, 350mW) and CC2420 as its transceiver. Because the proposed encryption algorithm neither changes the size of JPEG 2000 code stream, nor interferes with the image transmission, we only care about the energy consumption in the processor. The energy consumption model is thereby simplified as:

$$E = t * P \tag{2}$$

where t is the time consumption, and P is the power of the processor.

The energy consumption results of JPEG 2000 coding with and without encryption are listed in Table 3. It is clear that the encryption does increase negligible computational overhead on energy consumption.

Table 3. Energy consumption comparison on Lena

Image	E (mJ)
JPEG 2000 coded image	229
JPEG 2000 coded image with encryption	294

5 Conclusion

In this paper, we proposed a fast lightweight image encryption algorithm in wireless multimedia sensor networks (WMSN) for protecting JPEG 2000 images. The encryption algorithm combines the ideas of selective encryption and joint compression and encryption. At the entropy coding stage of JPEG 2000 compression, the lookup table of probability model in MQ coder is disturbed to change the values of Q_e in secure. In this manner, only a tiny portion of data is needed to be encrypted to protect the whole image in WMSN. Since the size of the lookup table is fixed, the encryption efficiency of the proposed algorithm is better than many existing schemes where the data to be encrypted are proportional to image size. The security and performance of the proposed algorithm are analyzed and related experimental results are given, both of them indicate that the proposed algorithm is fast and secure. Furthermore, the compression performance of JPEG 2000 standard almost remains intact.

Acknowledgments. The work in this paper was supported by the National Natural Science Foundation of China (No. 61103211) and the Program for New Century Excellent Talents in University (No. NCET-12-0589).

References

1. Akyildiz, I., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Computer Networks 38(4), 393–422 (2002)
2. Akyildiz, I., Melodia, T., Chowdhury, K.R.: A survey on wireless multimedia sensor networks. Computer Networks 51(4), 921–960 (2007)
3. Cheng, H., Li, X.: Partial encryption of compressed images and videos. IEEE Transactions on Signal Processing 48(8), 2439–2451 (2000)
4. Podesser, M., Schmidt, H.P., Uhl, A.: Selective bitplane encryption for secure transmission of image data in mobile environments. In: IEEE Nordic Signal Processing Symposium (NORSIG 2002), Tromso-Trondheim, Norway (2002)
5. Sadourny, Y., Conan, V.: A proposal for supporting selective encryption in JPSEC. IEEE Transactions on Consumer Electronics 49(4), 846–849 (2003)
6. Pfarrhofer, R., Uhl, A.: Selective image encryption using JBIG. In: Dittmann, J., Katzenbeisser, S., Uhl, A. (eds.) CMS 2005. LNCS, vol. 3677, pp. 98–107. Springer, Heidelberg (2005)
7. Grangetto, M., Magli, E., Olmo, G.: Multimedia selective encryption by means of randomized arithmetic coding. IEEE Transactions on Multimedia 8(5), 905–917 (2006)
8. Liu, J.L.: Efficient selective encryption for JPEG 2000 images using private initial table. Pattern Recognition 39(8), 1509–1517 (2006)
9. Christopoulos, C., Skodras, A., Ebrahimi, T.: The JPEG2000 still image coding system: an overview. IEEE Transactions on Consumer Electronics 46(4), 1103–1127 (2000)
10. Chang, H.K.C., Liu, J.L.: A linear quadtree compression scheme for image encryption. Signal Processing: Image Communication 10(4), 279–290 (1997)
11. Lian, S., Sun, J., Wang, Z.: Perceptual cryptography on SPIHT compressed images or videos. In: IEEE International Conference on Multimedia and Expo (ICME 2004), Taipei, Taiwan, pp. 2195–2198 (2004)
12. Lian, S., Sun, J., Wang, Z.: A secure 3D-SPIHT codec. In: European Signal Processing Conference (EUSIPCO 2004), Vienna, Austria, pp. 813–816 (2004)
13. Wu, C.P., Kuo, C.C.J.: Design of integrated multimedia compression and encryption systems. IEEE Transactions on Multimedia 7(5), 828–839 (2005)
14. Martin, K., Lukac, R., Plataniotis, K.N.: Efficient encryption of wavelet-based coded color images. Pattern Recognition 38(7), 1111–1115 (2005)
15. Wen, J., Kim, H., Villasenor, J.D.: Binary arithmetic coding with key-based interval splitting. IEEE Signal Processing Letters 13(2), 69–72 (2006)
16. Kim, H., Wen, J., Villasenor, J.D.: Secure arithmetic coding. IEEE Transactions on Signal Processing 55(5), 2263–2272 (2007)
17. Martin, K., Member, S., Plataniotis, K.N.: Privacy protected surveillance using secure visual object coding. IEEE Transactions on Circuits and Systems for Video Technology 18(8), 1152–1162 (2008)
18. Li, H., Zhang, J.: A secure and efficient entropy coding based on arithmetic coding. Communications in Nonlinear Science and Numerical Simulation 14(12), 4304–4318 (2009)
19. Taneja, N., Raman, B., Gupta, I.: Partial encryption on SPIHT compressed images. In: Chaudhury, S., Mitra, S., Murthy, C.A., Sastry, P.S., Pal, S.K. (eds.) PReMI 2009. LNCS, vol. 5909, pp. 426–431. Springer, Heidelberg (2009)
20. Hermassi, H., Rhouma, R., Belghith, S.: Joint compression and encryption using chaotically mutated Huffman trees. Communications in Nonlinear Science and Numerical Simulation 15(10), 2987–2999 (2010)

21. Wong, K.W., Lin, Q., Chen, J.: Simultaneous arithmetic coding and encryption using chaotic maps. IEEE Transactions on Circuits and Systems Part II: Express Briefs 57(2), 146–150 (2010)
22. Wang, W., Peng, D., Wang, H., Sharif, H.: A cross layer resource allocation scheme for secure image delivery in wireless sensor networks. In: 2007 International Conference on Wireless Communications and Mobile Computing (IWCMC 2007), Hawaii, USA, pp. 152–157 (2007)
23. Wang, W., Peng, D., Wang, H., Sharif, H., Chen, H.H.: Energy-constrained quality optimization for secure image transmission in wireless sensor networks. Advances in Multimedia 1, 1–9 (2007)
24. Wang, W., Peng, D., Wang, H., Sharif, H.: An adaptive approach for image encryption and secure transmission over multirate wireless sensor networks. Wireless Communications and Mobile Computing 9, 383–393 (2009)
25. Wang, H., Hempel, M., Peng, D., Wang, W., Sharif, H., Chen, H.H.: Index-based selective audio encryption for wireless multimedia sensor networks. IEEE Transactions on Multimedia 12(3), 215–223 (2010)
26. Wang, H., Hempel, M., Peng, D., Wang, W., Sharif, H., Chen, H.H.: On energy efficient encryption for video streaming in wireless sensor networks. IEEE Transactions on Multimedia 12(5), 417–426 (2010)
27. Taubman, D.S., Marcellin, M.W.: JPEG2000: Compression Fundamentals, Standards and Practice. Kluwer Academic Publishers (2002)
28. OMNeT++, `http://www.omnetpp.org/`

Evaluating Selective ARQ and Slotted Handshake Based Access in Real World Underwater Networks

Haining Mo[1], Lina Pu[1], Yibo Zhu[1], Zheng Peng[1],
Zaihan Jiang[2], and Jun-Hong Cui[1]

[1] Computer Science and Engineering Department,
University of Connecticut, Storrs, CT, USA
{haining.mo,lina.pu,yibo.zhu,zhengpeng,jcui}@engr.uconn.edu
[2] Acoustic Division, U.S. Naval Research Lab., Washington DC, USA

Abstract. Medium Access Control (MAC) is an essential component of protocol stacks in Underwater Acoustic Networks (UANs). Numerous dedicated UAN MAC protocols have been proposed and studied via analysis and simulations. However, limited work has been done on evaluating these protocols in real ocean environments. To achieve a better understanding on how MAC protocols perform in real world UANs, we implemented Selective ARQ and Slotted Handshake based Access (SASHA) on UAN nodes. SASHA embraces some most essential and representative techniques in UAN MAC design, including selective ARQ, time slotting, handshake and collision avoidance. Moreover, a sea test was conducted at Atlantic Ocean to evaluate the performance of SASHA. With the experimental data, we are able to study how the aforementioned techniques affect the performance of SASHA. we also analyze the hop-by-hop and end-to-end behavior of SASHA. Specifically, we investigate the transmission delay and queuing delay of a data packet on one hop. From the findings, some issues are discovered and the corresponding design guidelines are emerged.

1 Introduction

The last decade has witnessed a remarkable advance in Underwater Acoustic Networks (UANs) [1–3]. Despite the substantial progress in UAN protocol design, limited work has been done on protocol implementation and test in real world UANs, mainly due to the tremendous difficulties in real system implementation and conducting sea tests. Consequentially a complete picture of the behavior and performance of the protocols in real world UANs is missing. To advance towards filling this void, in this paper, we implemented a protocol SASHA, that incorporates some widely adopted techniques in the design of UAN MAC protocols. Beyond the implementation, we conducted sea tests to evaluate the performance of SASHA. By doing this, we hope to achieve a better understanding on how these widely employed techniques in UAN MAC design behave and perform in real world underwater networks.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 206–220, 2013.

UAN protocol design is challenging given the unique features of underwater environments including long propagation delays, high error probabilities and dynamic topologies. Beyond that, protocol implementation, test and evaluation in UANs have been associated with even further difficulties. Protocols need to be implemented on micro-controllers and control underwater acoustic modems such as Teledyne Benthos Modem [5]. To test and evaluate the implemented protocols, multiple UAN nodes, each of which weighs up to hundreds of pounds, have to be deployed. Based on our previous sea test experience, deploying an 8-node UAN can easily take up to one full day. Therefore, despite many dedicated UAN protocols that have been studied based on analysis and simulations, few of them have ever been tested or evaluated in sea experiments.

In this paper, we focus on Medium Access Control (MAC) for UANs. UAN MAC protocols can be coarsely classified into two categories: random access based ones and coordination based ones. Aloha-based protocols [6] usually employ a random channel access scheme. Random access based protocols introduce a minimum overhead from control packets but usually suffer more from collisions. By contrast, coordination based protocols generally initiate a dialogue among potential contenders prior to data transmission in order to avoid collisions. As a result, coordination based protocols incur a larger overhead but usually achieve a much smaller collision probability. In a UAN with a relatively high traffic load and node density, a random access based protocol can severely degrade the overall efficiency of data transmission due to the frequent occurrence of collisions. In this paper, we concentrate on coordination based MAC protocols.

In coordination based protocols, different approaches have been explored to compete for the channel, among which two-way handshake is the mostly used one. In [7], the authors utilized RTS/CTS based handshake. In [8], an improved handshake scheme was proposed. By deliberately delaying transmissions of CTS and data packets as well as utilizing the propagation delay gap, the proposed APCAP protocol was able to improve the overall efficiency of data transmissions. Besides the handshake mechanism, the authors in [9] used a tone-based approach to compete for the channel among multiple potentially contending UAN nodes.

Only handshake itself is usually not enough to completely eliminate collisions [10]. Therefore, in [7], collision avoidance and time slotting were employed, which allows packets to be sent out only at the beginning of time slots. Also packet train is critical to improve data transmission throughput and therefore was utilized by [7]. The authors in [9] utilized a random back-off scheme, which significantly lowers network traffic and therefore reduces collisions. Through analysis and simulations, the aforementioned approaches have been proved to be effective and therefore widely used by coordination based MAC protocols.

Great efforts have been devoted to simulation based analysis of these popular techniques, including handshake, time slotting, collision avoidance and back-off. However, simulations have two limitations. On one hand, most simulators have limited capability in reflecting the highly dynamic nature of the underwater environment and therefore there could be a nontrivial gap between the performance of a protocol in a simulator and that in real world sea tests. This was also supported by [11], which discussed the gap of this kind. On the other hand, there may be some facts unknown to the simulators that can only be revealed in real world sea tests. For instance, Pu et al. in [12] discovered that the long preamble of acoustic modems could significantly degrade the performance of UAN MAC protocols, which was not taken into consideration by simulators before.

Based on the above discussion, we understand that designing a MAC protocol involving the aforementioned techniques as well as implementing and testing it in real world UANs will be very helpful to study how general coordination based MAC protocols perform in real world. Petrioli et al. in [11] conducted a sea test to compare the performance of three UAN MAC protocols: CSMA, T-Lohi and DACAP. The tested protocols involved collision avoidance and back-off. However, other representative techniques in coordination based MAC protocols such as handshake, time slotting and selective ARQ were not involved. Besides, the scale of the deployed UAN was relatively small, consisting of only 3 nodes. The nodes were also located close to the shore since they had to be cabled to stations onshore for power supply and data transmission.

In this paper, we designed a MAC protocol called Selective ARQ and Slotted Handshake based Access (SASHA). As suggested by the name, SASHA is a coordination based protocol utilizing selective ARQ, time slotting, handshake and collision avoidance, which are all widely employed techniques in UAN MAC protocols. The reason why these techniques were chosen and a detailed description of them will be given in Section 2. Besides the design, we implemented SASHA on UAN nodes by utilizing Gumstix [4], Teledyne Benthos Modem [5] and a UAN protocol stack Aqua-NET [13]. Beyond that, we conducted a sea test to evaluate the performance of SASHA. A 9-node UAN was deployed 120 km off New Jersey shore between September 6th and 10th, 2012. The details of the sea test will be presented in Section 3. We tested SASHA with varying network sizes and traffic rates. Both the hop-by-hop and end-to-end performance of SASHA were measured and analyzed, as will be discussed in Section 4.

2 Protocol Description

SASHA is a coordination based MAC protocol, which employs handshake, time slotting, selective ARQ and collision avoidance. In this section, first we will present the protocol overview. Followed by that, the 4 key components of SASHA will de discussed in detail.

2.1 SASHA Overview

The overall work flow of SASHA is illustrated in Fig. 1. Node i and $i+1$ are the sender and the receiver while node $i-1$ and $i+2$ are two bystanders that can overhear packet transmissions between node i and $i+1$. An RTS/CTS exchange lasting two time slots is initiated between node i and $i+1$ to establish a conversation. Node $i-1$ which overhears the RTS and node $i+2$ which overhears the CTS, will back-off. After that, in the beginning of the next time slot, node i sends out an HDR, followed by a DATA packet train of 3 packets, in this example. HDR also causes node $i-1$ to keep silent. Due to the channel erasure, only 2 DATA packets are received at node $i+1$. Therefore node $i+1$ sends out a NACK, causing node $i+2$ to keep silent. Upon receiving the NACK, node i sends out an HDR in the next time slot, followed by the retransmitted DATA packet 2. This retransmission continues until an ACK is received at node i. Note that the transmissions of the control packets and the first DATA packet in the packet train are initiated in the beginning of a time slot.

Besides DATA packets, there are 5 types of control packets in SASHA. RTS/CTS are used for the handshake procedure. HDR is a new type of control packet sent out prior to the transmission/retransmission of DATA packets. It carries the information on how many DATA packets will be sent out following the transmission of HDR. HDR servers the purpose of both selective ARQ and collision avoidance. First, it informs the receiver of the expected number of DATA packets and therefore the receiver is able to construct a NACK/ACK packet. Second, a bystander overhearing an HDR can estimate the duration of the coming data transmission session based on the information embedded in HDR. Therefore it is able to choose an appropriate back-off period. NACK and ACK packets are used for selective ARQ [14]. Although HDR seems to be enough for the purpose of collision avoidance, we let RTS/CTS/NACK carry similar information to HDR, which is the number of DATA packets in the coming transmission session. A bystander overhearing an RTS/CTS/NACK can thus back-off accordingly. The purpose of this decision is to add redundancy and therefore an HDR loss will not invalidate collision avoidance.

2.2 Handshake and Time Slotting

RTS/CTS based handshake is employed by most coordination based UAN MAC protocols to alleviate collisions. However, only handshake is not enough to eliminate collisions, especially in UANs where the propagation delay is no longer negligible compared to the packet transmission duration. In [7], a time slotting mechanism is employed to address this issue. RTS/CTS can only be sent out at the beginning of a time slot. The length of a time slot is set to be the maximum propagation delay plus the CTS transmission duration. SASHA employs the same handshake and time slotting mechanism.

2.3 Selective ARQ

In coordination based MAC protocols, a nontrivial overhead is incurred by the handshake procedure. Therefore, to improve channel utilization and energy efficiency, a packet train of multiple DATA packets is usually transmitted after a successful handshake. Most current implementations of UAN MAC protocols have not incorporated selective ARQ into this packet train scheme. After a successful handshake, only one DATA transmission session is allowed. This means that lost/unacked DATA packets will not be retransmitted immediately. Instead, the communication pair has to re-compete for the channel via a new handshake

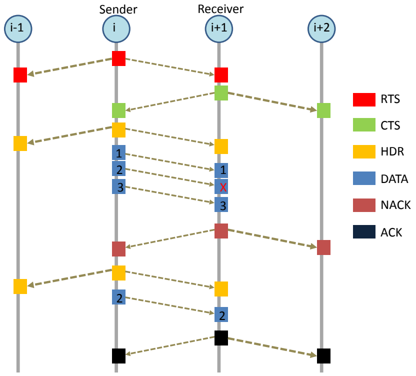


Fig. 1. SASHA overall work flow

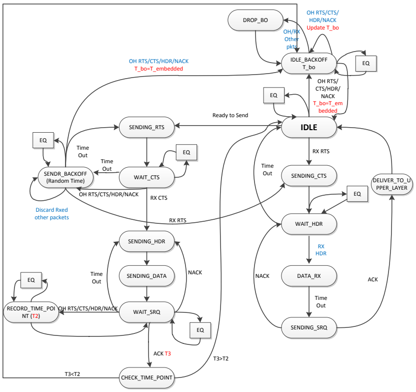


Fig. 2. SASHA state machine

procedure in order to initiate a retransmission. With this scheme, an ACK loss will lead to not only a retransmission but also a re-competition since the sender would assume that no DATA packet was received. If ACK loss happens frequently, a significant overhead from handshake will be imposed. We may assume that the probability of ACK loss is much smaller than that of DATA packets since an ACK is much shorter. However, channel asymmetry has been observed in [15], which discovered that in real world UANs, on a single channel, the backward link may have a much worse link quality than the forward link or vice versa. This is also supported by our previous field test experience, when ACK loss sometimes occurred with a comparable probability to DATA packet loss.

Motivated by the above discussion, SASHA implemented a selective ARQ scheme associated with the packet train mechanism. After a successful handshake, multiple consecutive DATA retransmissions are allowed until an ACK is received acknowledging all the DATA packets in the packet train. If a NACK is received, the sender will retransmit the lost DATA packets. If no NACK/ACK is received after time out, the sender will retransmit all the last transmitted DATA packets. Therefore with selective ARQ, one handshake can guarantee that the whole packet train can be received successfully. Selective ARQ effectively tackles the ACK loss issue and reduces the overhead brought by handshake.

2.4 Collision Avoidance

The purpose of collision avoidance is to eliminate collisions with ongoing DATA transmission sessions, which is achieved with the help of RTS, CTS, HDR and NACK packets.

RTS and CTS carry the length of the packet train. HDR contains the number of the following DATA packets. NACK embeds the number of the missing DATA packets. In a word, they are able to inform the overhearing nodes of the number of DATA packets to be transmitted next, noted as N. Upon overhearing an RTS, a node has to keep silent until transmissions of the coming CTS, HDR, DATA and ACK are completed. Therefore the number of time slots to back-off after overhearing an RTS is:

$$N_{bo\_rts} = 3 + \lceil \frac{(N * D_t + D_p)}{T} \rceil \tag{1}$$

The first 3 time slots stem from CTS, HDR and ACK, each of which takes one time slot, while the second part of the equation accounts for the transmission delay plus the propagation delay of DATA packets. Here D_t is the transmission delay of a DATA Packet and D_p is the maximum propagation delay of a packet. T is the length of a time slot, which is equal to the maximum propagation delay plus the transmission duration of a control packet.

Similarly, we can calculate the number of time slots to back-off upon overhearing a CTS, HDR and NACK (N is obviously different under different conditions):

$$N_{bo\_cts} = 2 + \lceil \frac{(N * D_t + D_p)}{T} \rceil \tag{2}$$

$$N_{bo\_hdr} = 1 + \lceil \frac{(N * D_t + D_p)}{T} \rceil \tag{3}$$

$$N_{bo\_nack} = 2 + \lceil \frac{(N * D_t + D_p)}{T} \rceil \tag{4}$$

2.5 SASHA State Machine

The state machine of SASHA is shown in Fig. 2. It is composed of two main threads: the sending thread and the receiving thread, which respectively reflects state transitions of the sending and receiving procedure.

Initially, a node is in the IDLE state. If there is one or multiple outgoing packets, it goes into the SENDING_RTS state, sends RTS and steps into the WAIT_CTS state. If no CTS is received, the node branches into state SENDER_BACKOFF and backs-off accordingly. Otherwise, it transits to state SENDING_HDR, sends HDR out followed by DATA packets and gets to state WAIT_SRQ. In state WAIT_SRQ, with timing out or reception of a NACK, the node goes back to state SENDING_HDR and initiates a retransmission. On the other hand, if an ACK is received, it goes to either state IDLE_BACKOFF or state IDLE depending on whether an additional back-off is required.

For the receiving thread, a received RTS propels a node into state SNEDING_CTS, where it replies a CTS and proceeds to state WAIT_HDR. After receiving an HDR, it progresses into the DATA_RX state to collect incoming DATA packets. Then the node goes to the SENDING_SRQ state and transmits a NACK or an ACK depending on the number of DATA packets received. The former one leads the node back to state WAIT_HDR, expecting a retransmission, while the latter one relieves the node back to the IDLE state. In the IDLE state, an overheard RTS/CTS/HDR/NACK forces the node into state IDLE_BACKOFF, where it keeps silent according to the information embedded in that very control packet.

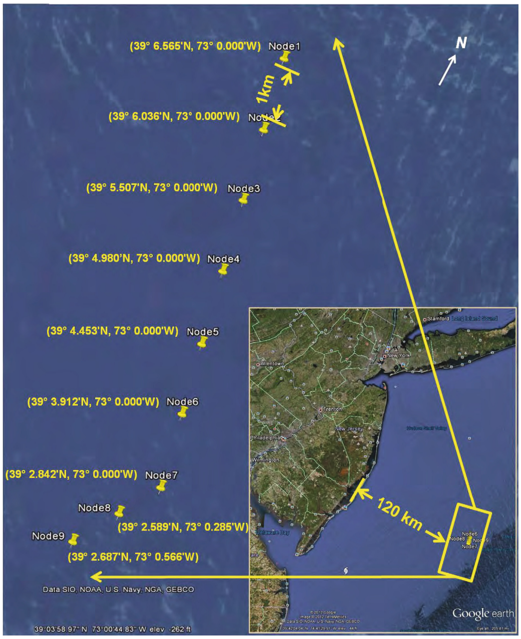

(39° 6.565'N, 73° 0.000'W) Node1
1km
(39° 6.036'N, 73° 0.000'W) Node2
N
(39° 5.507'N, 73° 0.000'W) Node3
(39° 4.980'N, 73° 0.000'W) Node4
(39° 4.453'N, 73° 0.000'W) Node5
(39° 3.912'N, 73° 0.000'W) Node6
(39° 2.842'N, 73° 0.000'W) Node7
Node8
(39° 2.589'N, 73° 0.285'W)
Node9
(39° 2.687'N, 73° 0.566'W)
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
39°03'58.97" N 73°00'44.83" W elev -262 ft
Stamford
Long Island Sound
New York
Allentown
Trenton
New Jersey
Hudson Shelf Valley
Philadelphia
Wilmington
Delaware Bay
120 km
Node6
Node8
Node9
Node7
© 2012 Google
Image © 2012 TerraMetrics
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
39°42'04.04" N 74°41'29.91" W elev 44 ft
Google earth
Eye alt 205.81 mi

Fig. 3. Sea test location and deployment

3 Sea Test Setting

To study the behavior and performance of SASHA, we conducted a sea test between September 6th and 10th, at Atlantic Ocean, 120 km off New Jersey shore. In this section, we are going to describe the network deployment and topology, hardware and software parameter settings as well as 3 successful tests that were done.

3.1 Network Deployment and Topology

During the sea test, we deployed 9 UAN nodes, at the locations as shown in Fig. 3. The nodes were located 120 km offshore at the depth around 80 m. The average height of the sea wave during the test was between 1.5 m and 2.5 m. The deployment area embraced wave, tide, salinity and temperature variance as well as fish and marine mammal movement. The average distance between two adjacent UAN nodes is 1 km while the distance between two end nodes is 7.3 km.

In the sea test, we focused on the string topology, mainly due to its popularity in a wide range of UAN applications. For example, it can be employed to send an image from the bottom of the ocean to a surface base station spanning several relay nodes. As shown in Fig. 3, our deployment provided a string network with up to 9 nodes. Also we chose to involve only some nodes in the middle to form networks with smaller sizes.

3.2 Hardware and Software Parameter Settings

In the sea test, each UAN node was equipped with one ATM-885 Teledyne Benthos modem, which operated at a frequency band between 16 kHz and 21 kHz. The available operation rates of Benthos modems range from 140 bps to 2400 bps. In this test, we adopted operation rates of 300 bps and 600 bps due to the consideration of both communication reliability and efficiency. Higher rates led to unacceptable packet loss rates while lower ones largely increased packet transmission delays.

For the software parameters, we experimented with different DATA packet lengths and traffic generation rates. 200 B and 400 B DATA packets were tested. Regarding the packet train length, we basically stuck with 2 since a packet train longer than 2 took a significantly long time for one round of communication due to the severe packet loss. The sender ran a Poisson Traffic Generator on the application layer in Aqua-NET. During the test, we experimented with the traffic generation rate equal to 0.005, 0.015 and 0.024.

3.3 Three Successful Tests

We managed to carry out 3 successful tests. The modem power level, modem operation rates and node counts of different tests are shown in Table 1.

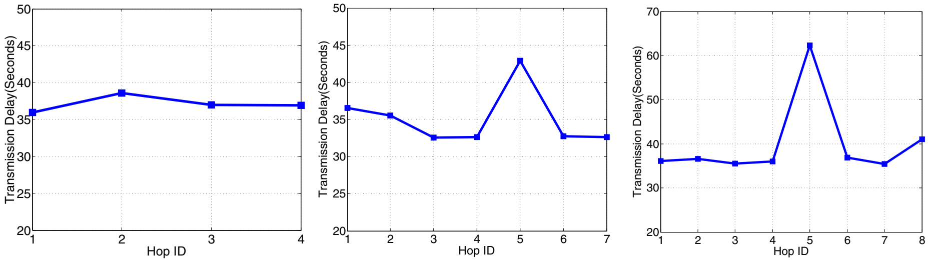
Table 1. Three successful tests

Test No.	Power Level	Operation Rate	Node Count	Packet Len	Train Len	Traffic Rate
1	1	300bps	5	100B	2	0.015
2	1	600bps	8	200B	2	0.005
3	1	300bps	9	200B	2	0.005

We formed 5-node, 8-node and 9-node (correspondingly 4-hop, 7-hop and 8-hop) networks. In all the three tests, we chose to set the transmission powers of Benthos modems to the lowest level in order to ensure that a node could reach only its immediate neighbors. The packet length and packet train length were both relatively small, considering the high probability of packet loss. The traffic generation rates were also selected to be low. Due to the significantly long end-to-end delays we observed during the tests, a high traffic generation rate could easily overwhelm the network.

4 Sea Test Results

We conducted sea tests to study the behavior and performance of SASHA in real world UANs. We are mostly interested into how SASHA behaves hop-by-hop wise. This mainly includes the packet delivery delay on a single hop and what factors contribute to this delay. The hop-by-hop behavior of SASHA fundamentally leads to its end-to-end performance. It provides insights on how



(a) 4-hop transmission delay (b) 7-hop transmission delay (c) 8-hop transmission delay

Fig. 4. Transmission delays of 4-hop, 7-hop and 8-hop networks

the techniques employed by SASHA perform in real world applications. In this section, we will discuss the hop-by-hop behavior of SASHA, followed by a presentation of its overall end-to-end performance, including the packet delivery delay, throughput and packet delivery ratio.

4.1 Hop-by-Hop Performance

In SASHA, a node essentially has three types of actions during its lifetime. If it has nothing queued in its incoming queue, it simply stays in the IDLE state. Otherwise, the node takes one of the two actions. If the node is informed of a potential collision or fails to complete a handshake due to RTS/CTS loss, the node backs-off. In this case, the packets are queued at the node. We call the delay related to this type of action queuing delay. On the other hand, if the node senses no conflicting activity and succeeds in completing a handshake, a packet train will be sent out. The delay before the entire packet train is successfully received at the receiver is called transmission delay. If we look back at Fig. 1, transmission delay refers to the duration between when RTS is sent out in a successful handshake and when ACK is finally received. Note that if due to RTS/CTS loss, a node is forced to attempt multiple handshakes, only the time related to the last RTS/CTS exchange is accounted into the transmission delay. The time consumed by previous handshake procedures is considered to be part of the queuing delay. The delivery delay of a packet on a single hop is the summation of the transmission delay and queuing delay. In the rest of this section, we will analyze the hop-by-hop transmission delay and queuing delay in detail.

Transmission Delay on One Hop. The average transmission delays of the 4-hop, 7-hop and 8-hop tests are shown in Fig. 4. Transmission delay generally consists of the transmission time and propagation time of RTS, CTS, HDR, DATA, ACK and possibly NACK and retransmitted DATA. As shown in Fig. 4(a), the transmission delays over different hops in the 4-hop test were pretty consistent. The reason is that during this test, DATA packet loss rarely happened and therefore few retransmissions were involved. Nevertheless, the small variance on different hops mainly originated from the propagation delay difference on different hops, which was caused by distance difference as well as temperature and salinity variance affecting the sound propagation speed.

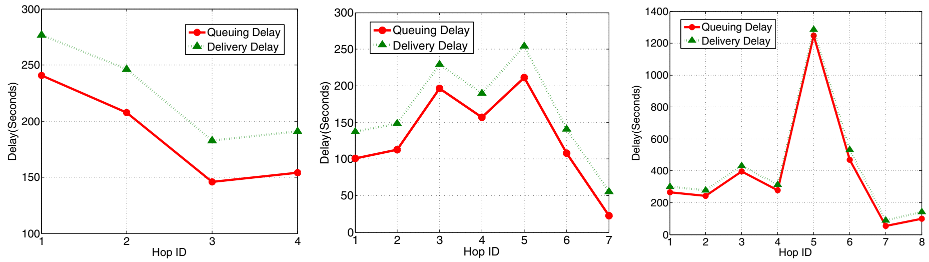
Similarly, in Fig. 4(b) and 4(c), transmission delays also maintained relatively stable values within the network expect that there was a peak point in both the 7-hop and 8-hop tests. From our sea test log file, we found that the much more significant transmission delay on Hop 5 in the 7-hop test, for instance, came from the retransmissions of DATA packets. On that very link, DATA packet loss occurred a lot and triggered the selective ARQ procedure and therefore a larger transmission delay was observed. Admittedly, selective ARQ increases the transmission delay for a packet train. However, it allows the entire packet train to be received with only one channel competition. Without selective ARQ, multiple competitions may be required to deliver the packet train, as discussed in Section 2.3. Due to the overhead from the handshake procedure, the overall per-hop delivery delay for the packet train would be significantly larger.

Therefore, the factor affecting transmission delay is the link quality of a given hop. On a hop with a poor link quality, the loss of DATA, NACK and ACK packets leads to retransmissions and therefore increases the transmission delay.

Queuing Delay on One Hop. Compared with the per-hop transmission delay, the queuing delay on one hop was much larger and accounted as the major part of the delivery delay, as shown in Fig. 5. Queuing delay of a DATA packet is defined to be the time from when the packet is received at a node to when the last RTS is sent out leading to a successful handshake. Earlier we mentioned that the queuing of a DATA packet is because of the back-off due to overhearing ongoing transmissions in the neighborhood or the failure of handshake incurred by RTS/CTS loss. During the tests, we discovered that another factor significantly contributing to the large queuing delay was the transmission range uncertainty which caused unexpected collisions.

In the 4-hop test, the largest queuing delay appeared on Hop 1. The reason is that we employed 0.015 as the traffic generation rate, which was the highest among all three tests. A traffic generation rate of 0.015 means that a DATA packet was generated at the source node averagely every 66 seconds. As we can see from Fig. 5(a), the average per-hop delivery delay in the 4-hop test was more than 150 seconds. Obviously, 0.015 was too aggressive. Under this circumstance, a lot of DATA packets were queued at the source node after being generated. This is why the first hop experienced the largest queuing and delivery delay. By contrast, the queuing delay reduced considerably on Hop 3 and 4 since the packet arrival rates on these two hops were much lower than that on the first hop. Another consequence of the high traffic generation rate was the increasing channel competition, which lowered the handshake success rate and therefore led to a large queuing delay. The underlying reason is that with a higher traffic generation rate, the source node tended to send out RTS at a higher frequency, which translated into higher RTS sending rates on succeeding nodes as well.

For the 7-hop test, the three middle hops, namely Hop 3, 4 and 5 experienced larger queuing as well as delivery delays than the other hops. The reason is



(a) 4-hop delays (b) 7-hop delays (c) 8-hop delays

Fig. 5. Queuing and delivery delays of 4-hop, 7-hop and 8-hop networks

that the middle nodes had more immediate neighbors than the edge nodes in the string topology. Therefore, they were likely to overhear more RTS/CTS from neighbors in both directions, which incurred larger back-off periods. Hop 7 had the lowest queuing delay because of two reasons. First, the traffic rate at the end of the network was much lower than those on the first several hops. Second, the sink node had only one immediate neighbor, which suffered from less competitions than the middle nodes. In this test, the traffic generation rate was 0.005, three times lower than that in the 4-hop test. This explains why the 7-hop test had overall lower queuing delays than the 4-hop test.

In the 8-hop test, we found an abnormally large queuing delay on hop 5, as shown in Fig. 5(c). There are already two known factors leading to this phenomenon. On one hand, the ongoing transmissions in the neighborhood caused both the sender and the receiver to back-off. On the other hand, Hop 5 suffered from a higher packet loss rate, as we discussed in Fig. 4. This could cause a nontrivial RTS/CTS loss, which incurred further back-offs on Hop 5.

However, after checking the test log files, we observed an extraordinary amount of handshake failures on Hop 5 due to RTS/CTS loss. This could not be explained by channel erasure only since RTS/CTS loss happened at a much higher rate than other control or DATA packets. After examination, we found the reason is that the two nodes on Hop 8 were able to reach the receiver on Hop 5. As a consequence, simultaneous transmissions of RTS on Hop 5 and 8 would cause collisions at Hop 5, which imposed back-offs. In another word, Hop 5 suffered from competitions with its immediate neighbors as well as with neighbors that were two hops away. With the frequent occurrence of this type of extra collisions, the queuing delay on Hop 5 became extremely long.

SASHA did not expect the above type of collision in a string network. The root reason is that even with very carefully selected modem transmission power levels, we still could not guarantee that a node was able and only able to communicate with its two immediate neighbors. This phenomenon is defined as transmission range uncertainty. In the above example, nodes on Hop 8 were able to talk to nodes on Hop 5. As a matter of fact, collisions caused by transmission range uncertainty also happened in the other two tests, but with much lower frequencies. Transmission range uncertainty was also observed and reported by [15]. It invalidates our assumption about network topology and might cause a much higher collision probability than expected during the handshake procedure in protocols like SASHA.

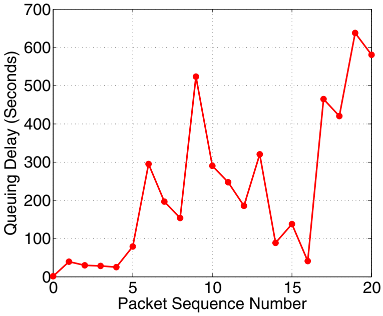
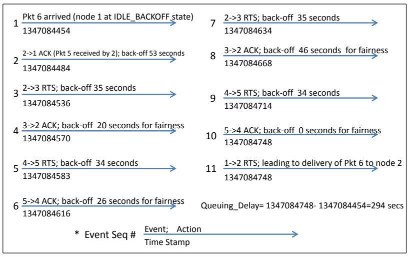


Fig. 6. Queuing delay on hop 1



1 Pkt 6 arrived (node 1 at IDLE_BACKOFF state)
1347084454

2 2->1 ACK (Pkt 5 received by 2); back-off 53 seconds
1347084484

3 2->3 RTS; back-off 35 seconds
1347084536

4 3->2 ACK; back-off 20 seconds for fairness
1347084570

5 4->5 RTS; back-off 34 seconds
1347084583

6 5->4 ACK; back-off 26 seconds for fairness
1347084616

7 2->3 RTS; back-off 35 seconds
1347084634

8 3->2 ACK; back-off 46 seconds for fairness
1347084668

9 4->5 RTS; back-off 34 seconds
1347084714

10 5->4 ACK; back-off 0 seconds for fairness
1347084748

11 1->2 RTS; leading to delivery of Pkt 6 to node 2
1347084748

Queuing_Delay= 1347084748- 1347084454=294 secs

* Event Seq # Event; Action
Time Stamp

Fig. 7. Queuing delay breakup for Packet 6

An Example of Per-Hop Queuing Delay. To better understand the factors contributing to the queuing delay, we take a closer look at one example. Fig. 6 shows the queuing delays of the first 21 DATA packets on node 1 in test 1. Since Packet 6 was the first DATA packet with a significant growth in queuing delay, we examine where this notable growth came from. Fig. 7 lists all the events that occurred from when Packet 6 was generated at node 1 to when RTS was sent out leading to the successful delivery of Packet 6 at node 2.

Basically there are three types of events contributing to the queuing delay of Packet 6 on node 1. First, when node 1 was in the BACK-OFF state and received/overheard an ACK, it did back-off for a random time period within a predefined range, as with event 2 and 4. This back-off is simply a design option in SASHA and serves the purpose of fairness, which means that a node has to perform a proactive back-off after one round of successful communication with the reception/overhearing of an ACK. This random back-off allows the surrounding nodes to have the same probability to start over and compete for the channel.

The second event associated with the queuing delay is overhearing ongoing transmissions in a node's immediate neighborhood. For instance, in event 3, node 1 overheard the RTS from node 2 to 3 and performed back-off. Similar situation happened in event 7. This type of back-off is expected and designed for collision avoidance in a node's immediate neighborhood.

Besides, transmission range uncertainty played a great role in composing the large queuing delay. As shown in event 5 and 9, node 1 overheard the transmission activity on hop 4, which caused unexpected back-off. This further verified our discussion on transmission range uncertainty in Section 4.1.

Discussion. In this section, we studied the transmission delay and queuing delay in one hop for SASHA. The transmission delay is mainly affected by the link quality on a given hop. The queuing delay usually stems from the back-off caused by collision avoidance and handshake failure. However, transmission range uncertainty can lead to an extraordinary amount of unexpected overhearing and collisions, which can significantly increase the queuing delay.

To alleviate the large queuing delay, we can work towards improving the fairness back-off scheme discussed in Section 4.1. The fairness back-off scheme lets a node temporarily give up the sending chance and therefore the neighboring node with the shortest remaining back-off time can be the first one to send next. This does enhance the fairness of SASHA but may lead to unnecessary back-offs. An optimal fairness back-off scheme should guarantee the fairness while eliminating unnecessary back-offs.

On the other hand, transmission range uncertainty contributes much more significantly to the large queuing delay than the other factors. One feasible solution might be to obtain the accurate network topology rather than assuming a network topology. To this end, a dynamic topology probing approach has to be in place.

4.2 End-to-End Performance

In this section, we investigate the end-to-end delivery delay, throughput and delivery ratio of SASHA. The overall end-to-end delivery delay of the 4-hop, 7-hop and 8-hop tests are shown in Fig. 8. As the network size grows larger, so does the end-to-end delivery delay. The significant growth in the delay for the 8-hop test stemmed from the unexpected collisions caused by transmission range uncertainty, as described in Section 4.1. The transmission range uncertainty problem becomes severer in larger networks, where more nodes lead to more overhearing and a larger collision probability.

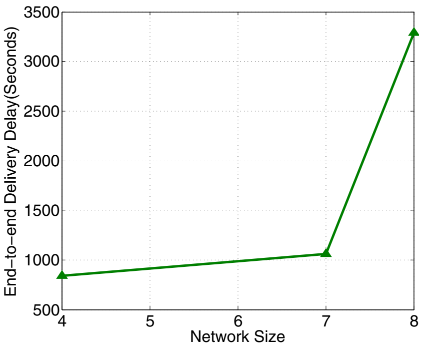


Fig. 8. End-to-end delivery delay

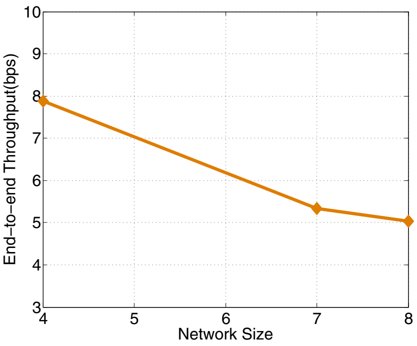


Fig. 9. End-to-end throughput

The end-to-end throughput decreases with the increase of the network size, as shown in Fig. 9. Also we can see that the achieved throughput in the 7-hop and 8-hop UANs were actually pretty low, which proved the difficulty of networked communication in real world underwater environments. In terms of end-to-end delivery ratio, all three tests achieved 100% delivery ratio. Since SASHA incorporates selective ARQ, it can guarantee the end-to-end reliability of each DATA packet.

5 Conclusions and Future Work

In this paper, towards gaining a better understanding on how UAN MAC protocols perform in real world underwater environments, we implemented SASHA on UAN nodes. SASHA employs some essential techniques in the design of coordination based UAN MAC protocols, including selective ARQ, time slotting, handshake and collision avoidance. In addition, a sea test was conducted at Atlantic Ocean to test SASHA. We analyze the performance of SASHA both hop-by-hop wise and end-to-end wise. Particularly, we investigate the transmission delay and queuing delay of a data packet on a single hop and what factors affect these two delays. Through investigation, we give some design guidelines to improve the performance of MAC protocols in real world UANs.

In terms of future work, we want to add a topology probing component to SASHA to make it aware of the accurate network topology. In this way, the large queuing delay with SASHA partly imposed by transmission range uncertainty is expected to be reduced. We also plan another sea test to compare different categories of UAN MAC protocols in the coming summer.

References

1. Cui, J.-H., Kong, J., Gerla, M., Zhou, S.: Challenges: building scalable mobile underwater wireless sensor networks for aquatic applications. IEEE Network, Special Issue on Wireless Sensor Networking 20(3), 12–18 (2006)
2. Akyildiz, I.F., Pompili, D., Melodia, T.: Underwater acoustic sensor networks: Research challenges. Ad Hoc Networks 3(3), 257–279 (2005)
3. Chitre, M., Shahabudeen, S., Stojanovic, M.: Underwater acoustic communicatin and networks: Recent advances and future challenges. Marine Technology Society Journal (1), 103–116 (2008)
4. Gumstix inc., `http://www.gumstix.com`
5. Benthos acoustic modem, `http://www.benthos.com`
6. Chirdchoo, N., Soh, W.-S., Chua, K.C.: Aloha-based MAC protocols with collision avoidance for underwater acoustic networks. In: Proceedings of IEEE INFOCOM (2007)
7. Molins, M., Stojanovic, M.: Slotted fama: a mac protocol for underwater acoustic networks. In: Proceedings of MTS/IEEE OCEANS (2006)
8. Guo, X., Frater, M.R., Ryan, M.J.: Design of a propagation-delay-tolerant MAC protocol for underwater acoustic sensor networks. IEEE Journal of Oceanic Engineering (2009)

9. Syed, A.A., Ye, W., Heidemann, J.: T-Lohi: A new class of MAC protocols for underwater acoustic sensor networks. In: Proceedings of IEEE INFOCOM (2008)
10. Fullmer, C.L., Garcia-Luna-Aceves, J.J.: Floor acquisition multiple access (FAMA) in single-channel wireless networks. Journal of Mobile Networks and Applications (1999)
11. Petrioli, C., Petroccia, R., Potter, J.: Performance evaluation of underwater MAC protocols: From simulation to at-sea testing. In: Proceedings of MTS/IEEE OCEANS (May 2012)
12. Pu, L., Luo, Y., Zhu, Y., Khare, S., Wang, L., Liu, B.: Impact of real modem characteristics on practical underwater MAC design. In: Proceedings of MTS/IEEE OCEANS (May 2012)
13. Peng, Z., Zhou, Z., Cui, J.-H., Shi, Z.: Aqua-Net: An underwater sensor network architecture - design and implementation. In: Proceedings of MTS/IEEE OCEANS (2009)
14. Comroe, R.A., Costello, D.J.: ARQ schemes for data transmission in mobile radio systems. IEEE Journal on Selected Areas in Communications 2, 472–481 (1984)
15. Pu, L., Luo, Y., Mo, H., Peng, Z., Cui, J.-H., Jiang, Z.: Comparing uderwater MAC protocol in real world. In: UCONN CSE Technical Report: UbiNet-TR13-03 (2013)

ActiviTune: A Multi-stage System for Activity Recognition of Passive Entities from Ambient FM-Radio Signals

Shuyu Shi, Stephan Sigg, and Yusheng Ji

National Institute of Informatics, Tokyo, Japan
{shi-sy,sigg,kei}@nii.ac.jp

Abstract. The amplitude of a received RF-signal is affected by physical phenomena, such as reflection, refraction or scattering due to objects and individuals in the signal propagation path. Activities in the proximity of a receiver can thus induce a characteristic pattern on amplitude-based features. We investigate the use of the radio frequency channel to detect activities. ActiviTune, our passive device-free recognition system, implements a multi-stage classifier to recognise activities and situations in an indoor environment leveraging amplitude-based features of RF signals from an ambient FM radio source. Comparing with other RF-based approaches, ActiviTune has the advantage of neither installing a transmitter generating the signal nor equipping the monitored entities with any active component of the system. We experimentally demonstrate the distinction of two dynamic activities, 'walking', 'crawling', and three static activities, 'empty room', 'standing', 'lying' with an average true positive rate of over 80%.

Keywords: Activity recognition, ambient context, multi stage recognition.

1 Introduction

In activity recognition, the majority of frequently applied sensing technologies require prior installation and repeated calibration within a changing environment. Examples are sensors for motion detection such as acceleration-type sensors [2,16] as well as video or RFID-based systems [6]. Clearly, the installation effort might diminish by proper integration of sensors in everyday objects, such as clothing. This has the potential to seamlessly foster the distribution of sensing equipment but also implicates the potential for inaccurate placement, thus increasing the importance of frequent sensor calibration. An approach that mitigates the necessity to equip users and also the frequent recalibration of the system is infrastructure-mediated sensing, introduced by Patel et al. [8]. Human actions might induce characteristic patterns in electric systems. These are analysed and utilized as a source to establish environmental awareness. This approach, however, is restricted to indoor settings. On the other hand, there are ubiquitously available, accurate and rich sensor classes from which we can detect activities of individuals indoors and outdoors. These sensors do not have to attached to the sensed entities which further implicitly reduces the effort required for calibration.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 221–232, 2013.

The RF-transceiver of electronic equipment might constitute such kind of sensor. Nearly all contemporary electronic devices contain some kind of interface to the RF-channel. Furthermore, since the available wireless spectrum is scarce [7], we are always surrounded by some sort of electromagnetic waves transporting signals associated, for instance, to audio (FM, AM), video (DVB-T), Speech (GSM, UMTS) or data (HSDPA, Bluetooth, Wifi, ZigBee). We propose to use the RF-channel as a sensing source to detect activities of individuals by monitoring peculiarities of RF-based features induced by these very activities. The contributions of this paper are

1. a study of features from the RF-channel which are suitable for the recognition of activities
2. a 2-stage activity classification system (ActiviTune) which distinguishes between static and dynamic activities in order to classify 'empty room', 'standing', 'lying', 'crawling' and 'walking'.
3. a case study with three probands in which we classify the above activities with high accuracy

In this paper we focus on the detection of five distinct activities 'empty room', 'standing', 'lying', 'crawling' and 'walking'. The case study is conducted with an Universal Software Radio Peripheral[1] (USRP) Software Defined Radio (SDR) device placed in a typical indoor environment. Using a multi-stage classification method, we discriminate between these activities with an overall rate of more than 80%.

The rest of this paper is structured as follows. The related work and preliminary studies are detailed in section 2. Section 3 discusses the theoretical foundation and feasibility of our recognition system, and describes implementation of the passive transmitter-free system based on the RF-channel. Experimental results from a case study are presented in section 4. Finally, section 5 draws our conclusion and discusses some open issues for future work.

2 Related Work

The feasibility of RF-based detection of situations was first indicated in 2006 by Woyach et al. [19] who describe the difference in the variance of RSSI fluctuations conditioned on environmental situations. The trajectory of an object and geometry of the environment impacted the RSSI fluctuations. Fuelled by these results, several authors investigated the localisation of not actively transmitting objects by RF-transceivers as summarized until 2010 in [9]. Zhang et al. used a grid of 870MHz nodes to investigate the accuracy to locate moving objects by RSSI estimates [21]. They derived the location of a moving object with about 1m accuracy. Youssef et al. demonstrated the localisation of individuals at a median accuracy of 1.82 meters by 802.11b nodes, continuously transmitting packets [20]. For the localisation, a passive radio map was constructed offline for a Bayesian inference algorithm [13]. The standard deviation of the RSSI was more stable to changes in the environment but more sensitive to movement when using a classifier trained on data from previous experiments [5]. Wilson and Patwari utilized

[1] `http://www.ettus.com`

Radio Tomographic Imaging (RTI) on the two-way RSSI variance [17] or RSSI mean fluctuations [18] between nodes arranged in a rectangle surrounding the monitored area for robust localisation of up to two individuals simultaneously. They reported that the variance of the RSSI can be used as an indicator of motion regardless of the average path loss that occurs due to dense walls and stationary objects. They experienced approximate errors of 0.9 meters and 0.45 meters for a moving and stationary individual respectively.

Related studies are conducted by Anderson et al. [1] and Sohn et al. [15] who have implemented activity recognition systems utilizing the fluctuation in GSM signal strength. Sohn et al. extracted seven features to classify a set of GSM measurements such as stationary, walking or driving with 85% accuracy. Recently, Ding et al. utilized the RF-noise emitted by electric components in contemporary automobiles to classify the traffic situation in front of a traffic light [4]. The authors achieved an accuracy of more than 0.95 in most of the cases with a single SDR employed using the Mean, Standard Deviation, Root of the Mean Squared (RMS) and Fast Fourier Transform (FFT) amplitude of the signal received on a frequency range of 2.4GHz.

The first study to report activity detection from passive entities was presented in 2011 by Scholz et. al [12]. The authors describe a system to detect walking, talking on a mobile phone and the state of the door in a typical office room with two SDR nodes (transmitter and receiver) at 900MHz placed on both sides of the door. Walking was detected by the number of peak-to-peak amplitudes greater than a trained threshold. The door context was triggered by a static change of amplitude in the signal. In order to detect a phone call a predefined area of the frequency spectrum was searched for a significant signal peak. The reported accuracy of the algorithm was on average above 80% for walking and above 90% for the other contexts. Sigg et. al considered the detection of the activities ’sitting’, ’walking’ and ’standing’ in a similar setting based on the RMS, Signal-to-Noise Ratio (SNR) and Average Magnitude Squared (AMS) of the Received Signal Strength Indicator (RSSI) [11]. The authors installed two or three SDRs, from which one was used to transmit a continuous signal at 900MHz or 2.4GHz in a room. The reported accuracy for these experiments was above 60%.

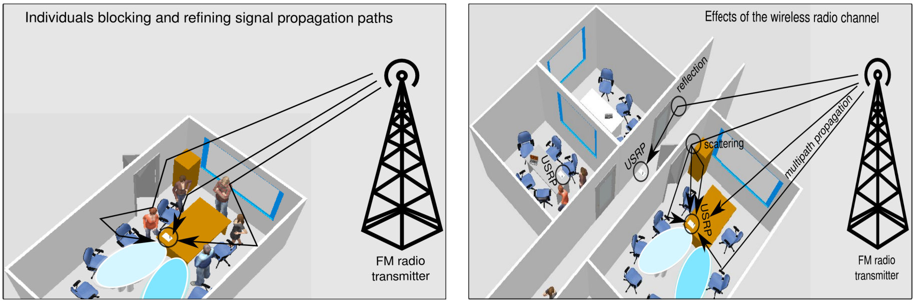
To the best of our knowledge, there are no results yet regarding the detection of activity of a passive, not actively transmitting entity from RF-channel measurements using environmental signals employed by a third party. All studies that utilize the RF-signal for activity detection require an active transmitter as part of the measurement system. In this case, the signal strength of the transmitted signal at the receiver is by several orders of magnitude higher and therefore easier to analyse. However, since the free wireless spectrum is scarce [7], we can expect to observe several non-noise signals at arbitrary location. Popular examples are FM- or AM-radio, DVB-T, WiFi, wimax, GSM or UMTS. In this paper, we consider the detection of activities from non-actively transmitting entities, utilising only signals from a third party, not under the control of the classification system.

3 RF-Based Activity Recognition System

In this section, we introduce some fundamental background on electromagnetic waves, which provides the theoretical basis for the realization of our passive transmitter-free system. Then, we will describe the system implementation, mainly consisting of a Raw Data Collector, a First Stage Classifier and a Second Stage Classifier.

3.1 Feasibility of RF-Based Activity Detection

Radio waves are electromagnetic waves, defined by their amplitude, phase and frequency. During signal propagation, radio waves are impacted by physical effects, for instance, damping, reflection and scattering. Assume a signal $m(t)$ at some frequency f_c [Hz]. Naturally, as signal propagation is roughly omnidirectional, the energy transmitted by a sender is propagated equally in all directions. In the event that a radio wave encounters any concrete structure such as an object or individual, the main signal component will be damped (continue its path with reduced energy) or even completely blocked. Additionally, the signal is typically reflected or scattered at this occasion. Reflection expresses the event that the signal wave bounces away from an object in a modified direction. Typically, the signal wave will also experience scattering, which is the splitting of a signal wave due to the not perfectly even structure of a hit object and the propagation of these signal waves into diverse directions. Figure 1 illustrates these effects for a signal emitted from a FM-radio station in an indoor situation.


Individuals blocking and refining signal propagation paths
FM radio transmitter
Effects of the wireless radio channel
reflection
USRP
scattering
multipath propagation
USRP
USRP
FM radio transmitter

(a) Changes to the signal propagation paths in an altered environment

(b) Effects of the mobile radio channel

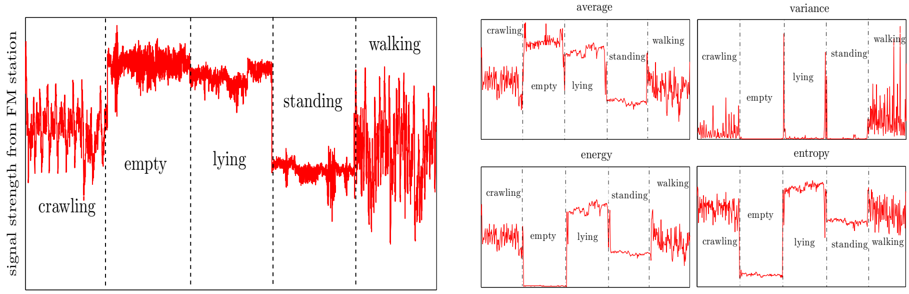
Fig. 1. Effects on the RF-signal during wireless signal transmission

As mentioned above, objects encountered by the electromagnetic waves during signal propagation can result in the alteration of physical effects experienced by signal components. This might then alter the properties of the signal at a receiver. However, various RF signals in different spectral bands, such as FM radio station, WiFi access points and GSM/UMTS/LTE base stations can all be leveraged as sources. For activity recognition, we believe that FM radio is currently best suited for RF-based systems for the following reasons. As shown in [3,10], leveraging broadcasted FM radio signals can augment the accuracy of indoor localisation, due to signal characteristics of FM radio, such as the low operating frequency range, the simple modulation mechanism and the wide area of coverage. It was demonstrated that these characteristics can be exploited to design more robust and discriminative signatures for RF-fingerprinting than WiFi [3,10] and GSM [10]. FM radio signals experience, when compared with WiFi, 3G or 4G signals, lower variation in signal strength over time [3]. Since we employ signal strength variation also for our features to distinguish activities, FM radio signals induce a lower process noise than signals from WiFi, 3G or 4G systems. For passive DFAR systems utilising ambient signal sources, also the sensitivity to changing weather conditions must be considered in the choice of alternatives. FM radio is, compared to the other named systems, which operate at higher frequencies, less susceptible to weather conditions, such as rain and fog. Additionally, in order to increase spectrum efficiency, spread spectrum techniques such as frequency hopping or code divisioning are employed in WiFi, 3G and 4G access points. In particular, participating devices that follow these schemes are required to closely comply to the parameters provided by the base station. Hopping schemes or CDMA codes are typically not disclosed to external devices. Since we assume that the ambient signal source is not under the control of the activity recognition system, it is natural to assume that the receiver is not capable of following any of these spread spectrum schemes. This likely results in significant variation in the observed signal strength or also in very low signal strengths in a given frequency band. For activity recognition, these spread spectrum techniques therefore increase process noise which makes a successful utilisation of the corresponding signals for passive DFAR systems more challenging. The modulation applied to FM radio is much simpler so that it is easier for a third party receiver to monitor the evolution of the signal. Furthermore, FM radio stations are widely implemented and continuously broadcast signals with higher coverage than WiFi, 3G or 4G systems. A passive DFAR system utilising FM radio can therefore be deployed virtually anywhere without considering the presence of surrounding implicit RF radio sources. Finally, FM radio is embedded in many contemporary electronic devices. For the above reasons, we believe that FM radio is best suited for the utilisation in a passive DFAR system.

We collected the amplitude of FM signals while five activities were conducted in proximity(cf. Fig 4). Figure 2 shows the received signal strengths for the five activities of a single individual over 3 minutes and the respective feature values at a two-second scale(feature extraction methods detail in equations (1)(2)(8) and (9)). We observe that different activities induce different patterns in the amplitude plots. Hence, we believe that, suitable features provided, it is feasible to correctly classify and recognize the activities based on the amplitude of FM-radio signals.

3.2 System Implementation

Our activity recognition system is passive and transmitter-free since monitored entities are not equipped with any part of the recognition system (passive) and the signal utilized originates from an ambient FM source, not under the control of the system. As summarized in figure 3, the system consists of three modules: Raw Data Collector, First Stage Classifier and Second Stage Classifier. The Raw Data Collector utilizes an USRP Software Defined Radio (SDR) device, tuned to a FM radio channel. In previous work [14], we have shown that a multi-stage recognition is superior to a one-stage recognition method regarding the classification accuracy. Therefore, the system exploits a two stage recognition method, which can firstly distinguish between 'stationary' and



(a) evolution of signal strength for all activities considered performed by a single subject

(b) Features for 5 activities utilising the signal strengths shown in figure 2a at a two-second scale

Fig. 2. Effects on the RF-signal during wireless signal transmission

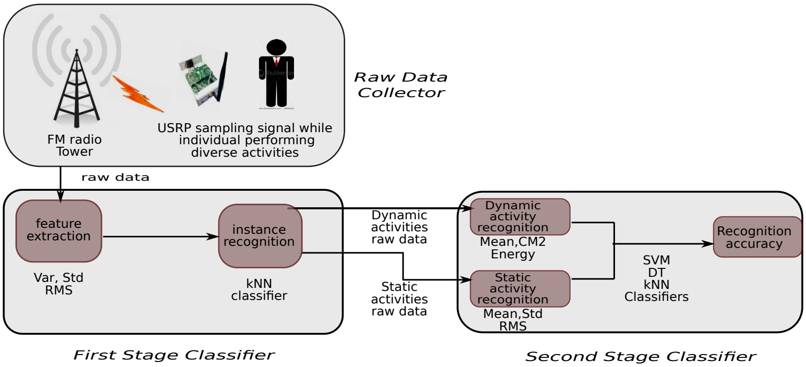


Fig. 3. An overview of the passive transmitter-free activity classification system

’dynamic’ activities based on fluctuation in the signal amplitude and then further distinguish activities in the set ’standing’, ’lying’, ’empty room’, ’walking’ and ’crawling’.

Raw Data Collector. The task of the Raw Data Collector is to convert the analog signals into digital signals and prepare these as the input of the following module. The amplitude of FM Radio signals transmitted by the FM Radio Station is measured by sampling the analog signals from the USRP front end. The system utilizes the WBX[2] daughter board with a USRP N210[3] SDR device. The antenna utilized is the VERT900[4] model with 3dBi antenna gain. The USRP SDR employs a general purpose software configurable FPGA-based digitizer equipped with a 12-bit dual-channel Analog-Digital Converter (ADC). We configure the daughter board to harmonize with one FM radio

[2] https://www.ettus.com/product/details/WBX

[3] https://www.ettus.com/content/files/2987_Ettus_N200-210_DS_FINAL_1.27.12.pdf

[4] https://www.ettus.com/product/details/VERT900

channel and the digitizer to successively sample the amplitude of its analog signals at a rate of T. The amplitude of sampled discrete signals from the device is expressed by $\zeta_{\text{rec}}(t)$ at the distinct time intervals $t = 1, 2, 3, \ldots, T$ within one second. Finally, to reduce the noise of the signal, the average amplitude is processed within a window size of T_1 so that $\frac{T}{T_1}$ amplitude values are collected per second.

First Stage Classifier. The activities classified can be categorized into 'dynamic', which covers 'walking' and 'crawling', and 'stationary' which is comprised of 'standing', 'lying' and 'empty'. We implement a hierarchical classification process consisting of two consecutive stages. In first stage classification, we attempt to discriminate between 'dynamic' and 'stationary' activities. Then, the actual activities are identified with the Second Stage Classifier.

First Stage Feature Extraction. Utilizing the USRP SDR device, we aggregate raw data containing various characteristics induced by specific activities. However, we experienced low accuracy utilising a direct classification of activities from raw data. We extract three features with a fixed window size of T_2 for recognition. A multitude of possible feature combinations have been considered to experimentally achieve best classification accuracy with a combination of Variance (Var), Standard Deviation (Std) and Root Mean Square (RMS) of the signal amplitude based on the experimental results of a case study in [14][5]. These features are defined as

$$\text{Mean} = \frac{\sum_{t=1}^{T_2} \zeta_{\text{rec}}(t)}{T_2} \tag{1}$$

$$\text{Var} = \frac{\sum_{t=1}^{T_2} \left(\zeta_{\text{rec}}(t) - \text{Mean}\right)^2}{T_2} \tag{2}$$

$$\text{Std} = \sqrt{\frac{\sum_{t=1}^{T_2} \left(\zeta_{\text{rec}}(t) - \text{Mean}\right)^2}{T_2 - 1}} \tag{3}$$

$$\text{RMS} = \sqrt{\frac{\sum_{t=1}^{T_2} (|\zeta_{\text{rec}}(t)| - \text{Mean})^2}{T_2}} \tag{4}$$

First Stage Instance Classification. To classify instances (stationary or dynamic) from the extracted features, the system utilizes a k-Nearest-Neighbor classifier (k-NN) with the default setting of the Orange data mining Toolkit[6]. We utilize a 10-fold cross validation approach to generate two data-subsets and only one is labelled and used to train the k-NN classifier while the other is used for classification.

[5] A condensed table summarising all classification accuracies achieved is available at `http://klab.nii.ac.jp/~shi/CAPS2012/table.pdf`

[6] `http://orange.biolab.si/`

Second Stage Classifier. Dependent on the outcome of the first stage classification, in the second stage, either the dynamic activities ’walking’ or ’crawling’, or the static activities ’standing’, ’lying’ or ’empty’ are distinguished. Again, we utilize a 10-fold cross validation to train the classifiers: in our case a Support Vector Machine (SVM), a Decision Tree (DT) and a k-Nearest-Neighbor (k-NN) approach each with the default configuration from the Orange data mining toolkit. Distinct feature combinations are utilized for static and dynamic activities.

Static Activity Recognition. In order to distinguish between static activities ’standing’, ’lying’ and ’empty’, we utilize in accordance to [14] the Mean, Std and RMS from the raw signal amplitude with a window size of T_2.

Dynamic Activity Recognition. In analogy to [14], one time domain feature of second cnetral moment (CM$_2$) and two frequency domain features of normalized Spectral Energy (Energy) and Entropy are leveraged for activity classification with a window size of T_2. The second central moment is defined as

$$\mathrm{CM}_2 = \mathrm{E}[\zeta_{\mathrm{rec}}(t) - \mathrm{E}(\zeta_{\mathrm{rec}}(t))]^2, \tag{5}$$

where E is the expectation. To compute the normalised spectral energy, the first step is to transform the amplitude from the time domain into the spectral values in frequency domain using the Fast Fourier Transform (FFT). The FFT can be represented as

$$\mathrm{FFT}(i) = \sum_{t=1}^{T_2} \zeta_{\mathrm{rec}}(t) e^{-j\frac{2\pi}{N}it}. \tag{6}$$

In equation (6) the value T_2 is the quantity of the samples in the FFT of a frame, in our case, we chose it equal to the window size T_2. The value $\mathrm{FFT}(i)$ denotes the i^{th} frequency component in this frame. The normalized energy of a signal can be computed as the squared sum of its probability density of spectrum in each frame. The probability of each spectral $\mathrm{FFT}(i)$ band can be computed as

$$\mathrm{P}(i) = \frac{\mathrm{FFT}(i)^2}{\sum_{j=1}^{T_2/2} \mathrm{FFT}(j)^2}. \tag{7}$$

Hence, we respectively calculate the normalized spectral energy and entropy as

$$\mathrm{Energy} = \sum_{i=1}^{T_2/2} \mathrm{P}(i)^2 \tag{8}$$

and

$$\mathrm{Entropy} = -\sum_{i=1}^{T_2/2} \mathrm{P}(i) \log_2 \mathrm{P}(i). \tag{9}$$

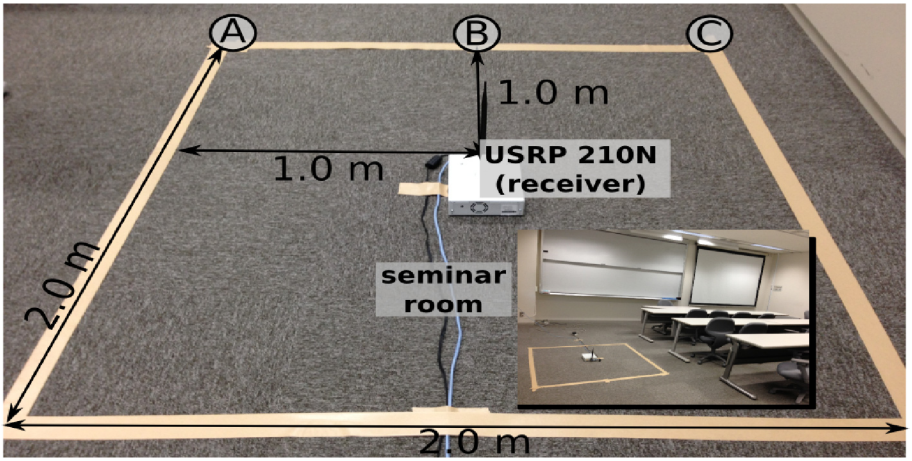


Fig. 4. Schematic illustration of the seminar room used in our case study

Recognition Accuracy. We calculate the true positive rate of the five activities as follows. For a set of k activities $\mathcal{A} = \{a_1, \ldots, a_k\}$ let $\mathcal{I}(a_i)$ be the total number of instances for activity a_i and $\mathcal{I}_{\text{cor}}(a_i)$ the number of correctly classified instances for this activity in which the classification matches the ground truth. We then define the accuracy by which an activity a_i could be detected as

$$\mathcal{ACC}(a_i) = \frac{\mathcal{I}_{\text{cor}}(a_i)}{\mathcal{I}(a_i)}. \tag{10}$$

4 A Case Study

To demonstrate the viability of the passive transmitter-free system, we performed a case study in a typical indoor environment.

4.1 Experimental Setup

The seminar room for the case study is depicted in figure 4. It is equipped with tables, chairs, a white board and some electronic equipment (a television, a projector, etc.). We used a $2m \times 2m$ square region as the recognition area with the USRP N210 in its center, configured to monitor the FM radio channel at 82.5MHz. In figure 4 we have marked the locations at which activities were performed with A, B and C (cf. table 1). Three subjects participated in the experiments. To collect sufficient data footage, every activity was conducted approximately 3 minutes by each subject with a sampling rate T of 256kHz. In our experiments, we found that $T_1 = 2^{12}$ and $T_2 = 32$ are most suitable to achieve favourable results so that 2 feature instances are fetched every second.

Table 1. Activities performed with respect to locations shown in figure 4

Activity	Location	Description
Stand	B	Standing still on location A
Lying	A-C	Lying across location A to C
Walk	All	Walking around freely in the 2 × 2 area
Crawl	All	Crawling around freely in the 2 × 2 area
Empty	–	Empty room without movement or individuals.

Table 2. Mean accuracy for the classification of the activities 'lying', 'standing', 'crawling', 'walking' and 'empty' from amplitude based features extracted from an FM-radio signal in the environment depicted in figure 4

(a) confusion matrix of accuracy of activities classified by Support Vector Machine algorithm

Ground-truth	Classified				
	empty	stand	lying	walk	crawl
empty	**.89(.053)**	.089(.021)	.007(.002)		.011(.002)
stand	.077(.019)	**.87(.057)**	.053(.017)		.004(.001)
lying	.053(.013)		**.94(.034)**		.008(.002)
walk	.073(.009)	.051(.011)	.051(.008)	**.74(.076)**	.080(.004)
crawl	.051(.011)	.005(.001)	.100(.039)		**.84(.063)**

(b) confusion matrix of accuracy of activities classified by Decision Tree algorithm

Ground-truth	Classified				
	empty	stand	lying	walk	crawl
empty	**.83(.102)**	.085(.036)	.070(.019)		.011(.005)
stand	.042(.007)	**.89(.031)**	.056(.009)		.004(.001)
lying	.034(.007)	.004(.001)	**.95(.010)**		.008(.003)
walk	.089(.026)	.073(.024)	.012(.003)	**.75(.133)**	.173(.082)
crawl	.037(.008)	.009(.002)	.107(.034)	.149(.022)	**.70(.110)**

(c) confusion matrix of accuracy of activities classified by k-Nearest Neighbor algorithm

Ground-truth	Classified				
	empty	stand	lying	walk	crawl
empty	**.92(.009)**	.037(.006)	.030(.004)	.004(.001)	.007(.001)
stand	.078(.014)	**.85(.077)**	.071(.031)		.004(.001)
lying	.049(.007)		**.94(.009)**		.008(.002)
walk	.034(.007)	.130(.029)	.051(.007)	**.73(.116)**	.051(.009)
crawl	.032(.003)	.005(.001)	.116(.041)	.163(.038)	**.68(.142)**

4.2 Results

The accuracy of the classification is presented in table 2. We ran a 10-fold-cross validation in the Second Stage Classifier. The tables show the mean accuracy for recognition of the five activities classified by the SVM, DT and k-NN classifiers respectively. Figures in brackets denote the standard deviation over the 10 test cases. Table fields with very low values (i.e. 0.0(0.0)) are left blank to improve readability.

We observe that all activities can be correctly recognized with a reasonably high average accuracy and low standard deviation. The SVM, DT and k-NN classifier achieve a mean accuracy of 0.86, 0.82 and 0.82. Also, based on the experimental results, we can conclude that the dynamic activities generally are less likely to be detected than the static ones. Remarkably, although the overall recognition accuracies achieved by various classifiers are similar, the precision for a given activity may differ. For instance, the accuracy for detecting 'crawling' can reach 84% utilizing the SVM classifier, however, it drops to 68% when recognized by the k-NN algorithm.

5 Conclusion and Future Work

The use of FM Radio for activity detection by recording and analysing the alteration of its wave attributes is a promising sensing approach. We have designed a passive transmitter-free activity recognition system and succeeded to classify five activities with an accuracy of over 80% with several classification algorithms, namely, Support Vector Machines, a Decision Tree and a k-Nearest-Neighbor approach. We conducted the experiments with three individuals for collecting raw data and adopted a two-stage approach to recognize these activities. Our experiments demonstrate that for the distinction between the activities 'lying', 'standing', 'empty room' 'walking' and 'crawling', the variation of the amplitude-based features is sufficient. In particular, the feature set of Var, Std and RMS is most suitable for the first stage recognition, and the combination of Mean, CM_2 and Energy can detect the dynamic activities best, while, another combination constituted by Mean, Std and RMS is chosen for static activities.

Still, some challenges remain and present future research questions for passive transmitter-free RF-based activity recognition systems. Among them are the investigation of the exact coverage of the sensor, to economize deployment as well as the feasibility of other radio signal frequency bands like UMTS or WiFi. Although, there still remain some open issues as mentioned above, we believe, with the constantly increasing spread of mobile device usage and the price advantage of FM radio, that RF sensing using FM radio signals is a promising and powerful source of contextual awareness for Ubiquitous Computing applications.

References

1. Anderson, I., Muller, H.: Context awareness via gsm signal strength fluctuation. In: 4th International Conference on Pervasive Computing, Late Breaking Results (2006)
2. Bao, L., Intille, S.: Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 1–17. Springer, Heidelberg (2004)

3. Chen, Y., Lymberopoulos, D., Liu, J., Priyantha, B.: Fm-based indoor localization. In: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, MobiSys 2012, pp. 169–182. ACM, New York (2012)
4. Ding, Y., Banitalebi, B., Miyaki, T., Beigl, M.: Rftraffic: Passive traffic awareness based on emitted rf noise from the vehicles. In: 2011 11th International Conference on ITS Telecommunications (ITST), pp. 393–398 (August 2011)
5. Kosba, A.E., Saeed, A., Youssef, M.: Rasid: A robust wlan device-free passive motion detection system. CoRR, abs/1105.6084 (2011)
6. Ogris, G., Lukowicz, P., Stiefmeier, T., Tröster, G.: Continuous activity recognition in a maintenance scenario: combining motion sensors and ultrasonic hands tracking. Pattern Analysis & Applications 15, 87–111 (2012)
7. Palaiyanur, H., Woyach, K., Tandra, R., Sahai, A.: Spectrum zoning as robust optimization. In: 2010 IEEE Symposium on New Frontiers in Dynamic Spectrum, pp. 1–12 (April 2010)
8. Patel, S.N., Robertson, T., Kientz, J.A., Reynolds, M.S., Abowd, G.D.: At the flick of a switch: Detecting and classifying unique electrical events on the residential power line. In: Krumm, J., Abowd, G.D., Seneviratne, A., Strang, T. (eds.) UbiComp 2007. LNCS, vol. 4717, pp. 271–288. Springer, Heidelberg (2007)
9. Patwari, N., Wilson, J.: Rf sensor networks for device-free localization: Measurements, models, and algorithms. Proceedings of the IEEE 98(11), 1961–1973 (2010)
10. Popleteev, A., Osmani, V., Mayora, O.: Investigation of indoor localization with ambient fm radio stations. In: 2012 IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 171–179 (March 2012)
11. Reschke, M., Starosta, J., Schwarzl, S., Sigg, S.: Situation awareness based on channel measurements. In: Proceedings of the Fourth Conference on Context Awareness for Proactive Systems, CAPS (2011)
12. Scholz, M., Sigg, S., Shihskova, D., von Zengen, G., Bagshik, G., Guenther, T., Beigl, M., Ji, Y.: Sensewaves: Radiowaves for context recognition. In: Video Proceedings of the 9th International Conference on Pervasive Computing, Pervasive 2011 (2011)
13. Seifeldin, M., Youssef, M.: Nuzzer: A large-scale device-free passive localization system for wireless environments. CoRR, abs/0908.0893 (2009)
14. Shi, S., Sigg, S., Ji, Y.: Activity recognition from radio frequency data: Multi-stage recognition and features. In: 2012 IEEE Vehicular Technology Conference, VTC Fall (2012)
15. Sohn, T., Varshavsky, A., LaMarca, A., Chen, M.Y., Choudhury, T., Smith, I., Consolvo, S., Hightower, J., Griswold, W.G., de Lara, E.: Mobility detection using everyday gsm traces. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 212–224. Springer, Heidelberg (2006)
16. Van Laerhoven, K., Gellersen, H.-W.: Spine versus porcupine: a study in distributed wearable activity recognition. In: Eighth International Symposium on Wearable Computers, ISWC 2004, vol. 1, pp. 142–149 (2004)
17. Wilson, J., Patwari, N.: Through-wall tracking using variance-based radio tomography networks. CoRR, abs/0909.5417 (2009)
18. Wilson, J., Patwari, N.: Radio tomographic imaging with wireless networks. IEEE Transactions on Mobile Computing 9, 621–632 (2010)
19. Woyach, K., Puccinelli, D., Haenggi, M.: Sensorless sensing in wireless networks: implementation and measurements. In: Proceedings of the Second International Workshop on Wireless Network Measurement, WiNMee (2006)
20. Youssef, M., Mah, M., Agrawala, A.: Challenges: Device-free passive localizsation for wireless environments. In: Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking (MobiCom 2007), pp. 222–229 (2007)
21. Zhang, D., Ni, L.: Dynamic clustering for tracking multiple transceiver-free objects. In: Proceedings of the 7th IEEE International Conference on Pervasive Computing and Communications, PerCom 2009 (2009)

From Decision Fusion to Localization in Radar Sensor Networks: A Game Theoretical View

Chuan Huang, Xu Chen, and Junshan Zhang

School of Electrical, Computer and Energy Engineering
Arizona State University, Tempe, AZ, 85201 USA
{huangch,xchen179,junshan.zhang}@asu.edu

Abstract. This paper considers the localization problem in a radar sensor network (RSN), where the estimation is made based on fusing the received signals from multiple radar sensors. For practical radar receivers, the moving target indication (MTI) technique is often adopted to suppress the clutter in the relatively small Doppler frequency shift regime, although it may filter out the desired target signal as well. As a result, when multiple radar sensors are deployed and the target is moving along one direction, it is likely that only a subset of the radar receivers can observe the target, which we call an *observation pattern*. In this paper, we explore how to utilize the information of all possible observation patterns to derive the Cramer-Rao lower bound (CRLB) for the localization problem, which is shown to hinge heavily on radars and the prior statistic information of the observation patterns. Next, we generalize the localization problem to the case for an area $\mathcal{S}$, and investigate the localization games between the RSN and the intrude target. We propose a two-stage Stakcelberg game framework to model the interactions between the RSN and the target, for cases that the target can adopt mixed and pure strategies, respectively. Finally, numerical results demonstrate that the proposed scheme can significantly improve the localization performance.

Keywords: Localization, radar sensor networks, observation pattern, radar range, Stackelberg game.

1 Introduction

Radar systems have been developed in the past decades, either to detect the existence of the moving targets or to estimate their parameters, including distance, direction, and speed. Conventional radar systems usually generate the decisions and estimations based on the observation at one radar receiver [1,2], e.g., the *monostatic radar*, with collocated transmitter and receiver, and the *bistatic radar*, with the transmitter and receiver located at different places. More recently, radar sensor networks (RSNs) are garnering much attention [3].

A critical issue of RSNs is the localization problem, i.e., estimating the position of the desired target by a sequence of temporally (or spatially) spanned observations. In particular, the observations could be the signal strength [4], range [2, 5] (i.e., the time delay of the received signals), angle of arrival [6], and so on, which can be individually or jointly explored to measure the location of the target by the RSN. In this paper, we focus on the case using the range information to locate the desired target.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 233–243, 2013.

In modern radar systems, the moving target indication (MTI) technique using a bandpass filter, has been widely adopted to suppress the clutter, which is mainly within the relatively small Doppler frequency shift (DFS) regime. Unfortunately, the MTI may also filter out the return from the desired target, and if this is the case it would make the target invisible to the corresponding radar receiver, leading to performance degradation of the radar receiver. To overcome this drawback incurred by the MTI, one plausible solution is to deploy multiple radar sensors in different locations of the monitored area to increase the diversity of observations. In [7], the authors studied the coverage issue for the RSN equipped with MTIs, for which a given point is said to be covered by the RSN if the target moving along arbitrary direction can always be *seen* by at least one of the radar receivers, i.e., the echos from the target cannot be filtered out by all receivers at the same time.

In this paper, we explore the cooperative localization of a high-speed target by using the RSN, assuming the MTI is used at each radar receiver. We first investigate the observation pattern for each point of interest in an area $\mathcal{S}$, which is defined as a vector with each element (with value 1 or 0) indicating whether the corresponding radar receiver can see the target or not. By exploiting the information of possible observation patterns at one point, the Cramer-Rao lower bound (CRLB) of the localization problem is derived, which is shown to be a joint function of the radar location and the statistics information of the observation pattern. We next extend to study the localization problem over the whole area $\mathcal{S}$ (with multiple points), and investigate the dynamic interaction between the RSN and the target. Given the fact that the RSN is deployed before the target intrudes, we propose a novel two-stage Stackelberg game framework, where the RSN is the leader to act first and the target is the follower to act subsequently. We also derive the Stackelberg equilibrium solutions to the localization games based on the principle of backward induction.

The remainder of the paper is organized as follows. Section 2 presents the system model and summarizes the main assumptions. Section 3 discusses the statistics of the received signals and also the observation patterns for the RSN. Section 4 investigates the CRLB of the localization problem. Section 5 investigates the localization game between the target and the RSN. Numerical results are presented in Section 6 to validate the theoretical analysis. Finally, Section 7 concludes this paper.

2 System Model

2.1 The RSN and the Target Model

In this paper, we consider a RSN consisting of K nomostatic radars, as shown in Fig. 1, which monitors a bounded area $\mathcal{S}$ and detects a possible intruding target. We denote the location of the i-th radar transceiver as $\boldsymbol{r}_i = (x_i, y_i)$, $i = 1, \cdots, K$[1]. Denote the point in area $\mathcal{S}$ as $\boldsymbol{s} = (x_s, y_s) \in \mathcal{S}$. We assume that the target moves with a constant speed V. To provide a worst-case study for the localization performance of the RSN, we assume that the target has all the information about the locations of the radar sensors of the RSN.

[1] For the case of three-dimension, the analysis is similar to this case, and thus omitted here for simplicity.

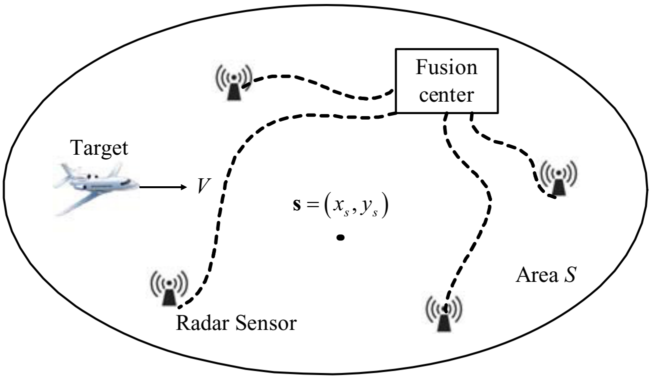


Fig. 1. An illustration of the radar sensor network, consisting of multiple radar sensors and monitoring an area $\mathcal{S}$, with a large number of points of interest $\{\boldsymbol{s}\}$

In practical systems, each radar transmitter periodically scans the whole monitoring area [1], in the sense that it sends a pulse along one spatial direction at one time, and based on the delayed echos, i.e., the range information, it estimates the location of the desired target.

2.2 DFS and MTI

The radar receivers make decisions based on the returns of the transmitted signals. Due to the movement of the objects in area $\mathcal{S}$, the received signal's frequency has been changed, i.e., there exists certain DFS, which is defined as the frequency difference between the emitted and the received signals due to the relative velocity between a radar receiver and the moving object. In particular, let f_s and f_i denote the frequencies of transmitted and the received signals at the i-th radar, respectively. The DFS, i.e., the difference between f_s and f_i, can be computed as [1]

$$\Delta f_i = f_i - f_s = \frac{2V \cos \varphi_i}{c} f_s, \tag{1}$$

where φ_i is the angle between the moving direction of the target and the line connecting the target and the i-th radar receiver, and c is the velocity of light. It is easy to see that for fixed φ_i, larger moving velocity V implies a larger DFS. In practice, the radar transmitted signal may be reflected by the desired high-speed target, as well as other low-speed objects, i.e., the birds and the mountains. Usually, the echos from the latter are called as the clutters, which are with much higher power than that from the desired target. However, since the DFS of the echo from the target is usually larger than that of the clutter, we can possibly distinguish the desired signal from both the additive noise and the clutter.

To suppress the clutter, the moving target indication (MTI) technique [1] is commonly adopted at the radar receiver before making the final decision. Essentially, the MTI applies a band-pass filter in the hope for that the filtered signal contains mainly the target signal when it is present. For ease of exposition, we assume that an ideal band-pass filter is adopted at each radar receiver, with the frequency response function given as

$$\mathcal{F}_i(f) = \begin{cases} 0, & |f| \leq f_i^c \\ 1, & |f| > f_i^c \end{cases}, \tag{2}$$

where f_i^c is the cut-off frequency of the band-pass filter at the i-th radar receiver. Based on (1) and the cut-off frequency f_i^c of the MTI, it follows that the target would be filtered out, when the angle φ_i satisfies the following condition

$$\varphi_0 \leq \varphi_i \leq \pi - \varphi_0, \tag{3}$$

where $\varphi_0 = \arccos\left(\frac{cf_i^c}{2Vf_s}\right)$.

2.3 Observation Patterns and Decision Fusion

Based on the analysis for the received signal at each radar receiver in the previous subsection, we know that each radar receiver cannot distinguish the cases that the target is not present and the target signal is filtered out by the MTI, which may decrease the probability of detection if the final decision is made only based on the observation at one radar receiver. However, when multiple radar sensors are deployed in area $\mathcal{S}$, at least a subset of them can "see" the desired target, i.e., the target signal cannot be filtered out by all the MTIs simultaneously. This indicates that there exists certain correlation among the observations at the radar receivers. First of all, we define an indication function I_i for the i-th radar sensor to describe the effect of the MTI as: $I_i = 0$, if $|\Delta f_i| \leq f_i^c$; $I_i = 1$, if $|\Delta f_i| > f_i^c$, for $i = 1, 2, \cdots, K$.

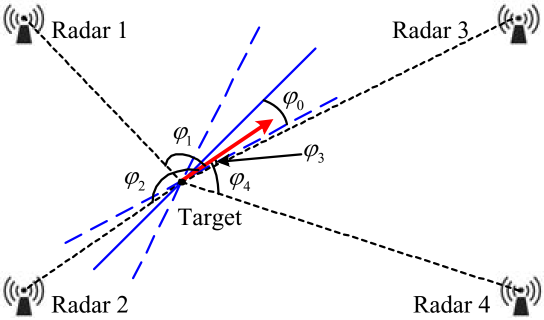


Fig. 2. One example for the observation pattern of the observed signals at the radar sensors

To better understand this phenomena, we first introduce an example to describe the relationship among the received signals at the radar receivers. As shown in Fig. 2, consider a target in the position $\boldsymbol{s}$, moving along a given direction (the arrow). We obtain four angle values φ_i, $i = 1, \cdots, 4$, as defined in (1) for the four radar receivers. For the considered case, it can be observed that only φ_1 is within the angle range given in (3), and thus the target signal will be filtered out by the MTI at the first radar receiver, while retained by the other three. Thus, the indication functions achieve the values $[I_1, I_2, I_3, I_4] = [0\ 1\ 1\ 1]$. We call the vector $[0\ 1\ 1\ 1]$ as one possible observation pattern associated to the point $\boldsymbol{s}$.

Remark 1. *It is easy to see that the observation patterns have the following two properties: 1)* location dependent*: The observation pattern is determined by the relatively spatial location between the target and the radar receivers; and 2)* moving direction dependent*: The observation pattern is also determined by the moving direction of the target.*

Based on the above analysis, when we change the moving direction of the target at one given point, we can obtain different observation patterns. We denote the set of all the possible observation patterns at the point $\boldsymbol{s}$ as $\mathcal{I}_s = \left\{\mathcal{I}_1^s, \mathcal{I}_2^s, \cdots, \mathcal{I}_{J_s}^s\right\}$, where $\mathcal{I}_j^s \in \mathcal{I}$, $j = 1, 2, \cdots, J_s$ and J_s is the total number of different observation patterns. Here $\mathcal{I}_s^j$ is defined as a vector of indication functions, i.e., $\mathcal{I}_s^j = [I_1, \cdots, I_K]$, which is within the following set

$$\mathcal{I} = \{[I_1, \cdots, I_K] : I_i \in \{0, 1\},\ i = 1, \cdots, K\}. \tag{4}$$

Based on the information of observation patterns, the RSN needs to first make decision on whether the desired target is present or not. This cooperative decision fusion scheme has already been discussed in [8], and the authors proposed an efficient linear cooperative detection scheme. From (3), we can see that the decision fusion needs to essentially utilize the angle information of the arrived signal.

3 CRLB for Localization

After the decision fusion, the RSN usually wants to localize and track the desired target online. In this paper, we investigate the localization problem for the RSN by utilizing the information of observation patterns. Since we further consider the information of observation patterns, we actually also explore the angle information of the received signal, i.e., the observation patterns, plus the range information to jointly complete the localization task, which is different from the range-only scheme discussed in [5].

Next, we derive the CRLB of the considered system under the given statistics of the observation patterns. Denote the probability that the observation pattern $\mathcal{I}_j^s$ happens as q_s^j, $j = 1, \cdots, J_s$, with $\sum_{j=1}^{J_s} q_s^j = 1$.

The distance between the considered location of the target $\boldsymbol{s}$ and the i-th radar receiver is given as [5]

$$l_i = 2|\boldsymbol{s} - \boldsymbol{r}_i|, \tag{5}$$

where $|\cdot|$ denotes the Euclidean norm operation, and $\boldsymbol{s} = (x_s, y_s)$ is unknown parameter to be estimated by the RSN. At the i-th radar receiver, it can extract one range measurement x_i, which is identical to the true distance plus a noise term, i.e.,

$$x_i = l_i I_{j,i} + w_i, \tag{6}$$

where $I_{j,i}$ is the i-th element of the j-th observation pattern $\mathcal{I}_j^s$. Note that if $I_{j,i} = 0$, the measure x_i only contains the noise term, which cannot bring positive contribution for the localization problem. Moreover, w_i is identically and independent distributed (i.i.d.) zero mean Gaussian random variables with variance σ_i^2, and σ_i^2 being a constant given by [5]

$$\sigma_i^2 = \frac{c^2}{\mu \alpha_i B^2}, \tag{7}$$

with μ is the transmit power, α_i is a system-dependent constant, and B represents the bandwidth of the probing signal.

Suppose $\widehat{\boldsymbol{s}}$ is one unbiased estimation of $\boldsymbol{s}$ based on the measurements $\{x_i\}$, and then the performance of this estimator is given by the following result.

Proposition 1. *The estimation error covariance matrix for any unbiased estimator* $\widehat{\boldsymbol{s}}$ *is lower bounded by*

$$\mathbb{E}\left\{(\widehat{\boldsymbol{s}} - \boldsymbol{s})(\widehat{\boldsymbol{s}} - \boldsymbol{s})^T\right\} \succeq \mathcal{J}^{-1}, \tag{8}$$

where $A \succeq B$ *means that* $A - B$ *is positive semidefinite and* $\mathcal{J}$ *is the nonsingular FIM as*

$$\mathcal{J} = \sum_{j=1}^{J_s} q_s^j \sum_{i=1}^{K} \mathbb{E}\left\{\bigtriangledown_{\boldsymbol{s}} \log p(x_i|\boldsymbol{s}, I_{j,i}) \bigtriangledown_s^T \log p(x_i|\boldsymbol{s}, I_{j,i})\right\}, \tag{9}$$

with $\bigtriangledown_s$ *denoting the gradient with respect to* $\boldsymbol{s}$*.*

The proof of the above proposition is similar to that in [5], by taking into account multiple observation patterns.

Since the noises w_i's are Gaussian, we obtain that

$$p(x_i|\boldsymbol{s}, I_{j,i}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{|x_i - l_i I_{j,i}|^2}{2\sigma_i^2}\right). \tag{10}$$

Thus, it follows that

$$\begin{aligned}\mathcal{J} &= \sum_{j=1}^{J_s} q_s^j \sum_{i=1}^{K} \mathbb{E}\left[|x_i - l_i I_{j,i}|^2\right] \frac{4}{\sigma_i^4} \frac{(\boldsymbol{s} - \boldsymbol{r}_i)(\boldsymbol{s} - \boldsymbol{r}_i)^T}{|\boldsymbol{s} - \boldsymbol{r}_i|^2} I_{j,i} &(11)\\ &= 4 \sum_{j=1}^{J_s} q_s^j \sum_{i=1}^{K} \frac{(\boldsymbol{s} - \boldsymbol{r}_i)(\boldsymbol{s} - \boldsymbol{r}_i)^T}{\sigma_i^2 |\boldsymbol{s} - \boldsymbol{r}_i|^2} I_{j,i}, &(12)\end{aligned}$$

which is a joint function of the radar location and the statistics information of the observation pattern.

Remark 2. *In this paper, we only consider the CRLB of the localization problem, which can apply to the general estimation schemes for the localization problem. Moreover, this bound is asymptotically achieved by some estimation schemes [5].*

Then, we define the mean square error (MSE) of the localization problem as

$$\gamma_s = \sum_{k=1}^{2} \lambda_k(\mathcal{J}^{-1}) = \operatorname{Tr}\left(\mathcal{J}^{-1}\right) \tag{13}$$

$$= \frac{\mathcal{J}_{1,1} + \mathcal{J}_{2,2}}{|\mathcal{J}|}, \tag{14}$$

where $\mathcal{J}_{1,1}$ and $\mathcal{J}_{2,2}$ are the elements at the entries $(1, 1)$ and $(2, 2)$ of the matrix $\mathcal{J}$, respectively. It is easy to obtain from (12) that

$$\mathcal{J}_{1,1} + \mathcal{J}_{2,2} = 4 \sum_{j=1}^{J_s} q_s^j \sum_{i=1}^{K} \frac{I_{j,i}}{\sigma_i^2}, \tag{15}$$

which is a constant.

4 Stackelberg Game for Localization

In this section, we study the localization problem over the given area $\mathcal{S}$. We introduce the localization game over area $\mathcal{S}$ taking into account the interplay between the RSN and the target. For the ease of exposition, we adopt the CRLB introduced in the previous section as the design metric for such game.

The motivation of considering localization game is that the goals of the RSN and the target are conflicting, i.e., the RSN would like to minimize the localization MSE, while the target wants to maximize it. In practical applications, the RSN is deployed first to monitor area $\mathcal{S}$, and the locations of the radar transceivers are fixed before the target intrudes. Therefore, we model the interaction between the RSN and the target as a two-stage Stackelberg game [9], in which the RSN is the leader to act first and the target is the follower to act subsequently.

From (13), we see that the available information of the target's intrude behavior will have a significant impact on the detection performance of the RSN. For example, if the target only chooses part of all the possible moving directions, not all the directions given in $\{\mathcal{I}_j^s\}$, there will be fewer observation patterns, which will change the MSE given in (13) and lead to a different localization performance of the RSN. This motivates us to study the Stackelberg game models with two different cases of target's strategy as follows.

1. Mixed Strategy: In this case, we consider that the target can move along different directions with certain probability. Specifically, we use a probability distribution $\boldsymbol{q}_s = \left[q_s^1, \cdots, q_s^{J_s}\right]$ over all J_s possible observation patterns to present the possibility of different moving directions of the target at point $\boldsymbol{s}$. Thus, the MSE $\gamma_{n,m}(\boldsymbol{s}, \boldsymbol{q}_{n,m})$ of the localization at the point $\boldsymbol{s}$ is given by (13).

We model the interaction between the RSN and the target as a Stackelberg game. Here RSN is the leader and chooses the locations of the radar transceivers before the target intrudes. The target is the follower and determines the optimal moving actions $\boldsymbol{q}_s$ for each point $\boldsymbol{s}$, given the locations of the radar transceivers $\{\boldsymbol{r}_i\}$ of the RSN.

By the principle of backward induction, given the locations of the radar transceivers $\{\boldsymbol{r}_i\}$ of the RSN, the target chooses the optimal action $\boldsymbol{q}_s$ to maximize γ_s over the whole area, i.e.,

$$\begin{aligned} &\max_{\boldsymbol{s}\in\mathcal{S}}\ \max_{\boldsymbol{q}_s}\ \gamma_s(\boldsymbol{r}_i,\boldsymbol{q}_s) \\ &\text{s.t.} \quad |\boldsymbol{q}_s| = 1, \forall \boldsymbol{s} \in \mathcal{S}. \end{aligned} \tag{16}$$

Then, the RSN would like to minimize the MSE of localization, i.e.,

$$\begin{aligned} \text{(P1)}\ &\min_{\{\boldsymbol{r}_i\}}\max_{\boldsymbol{s}\in\mathcal{S}}\max_{\boldsymbol{q}_s}\ \gamma_s(\boldsymbol{r}_i,\boldsymbol{q}_s) \\ \text{s.t.}\ &|\boldsymbol{q}_s| = 1,\ \forall \boldsymbol{s} \in \mathcal{S}. \end{aligned} \tag{17}$$

In general, Problem (P1) is non-convex [10], since $\gamma_s(\boldsymbol{r}_i,\boldsymbol{q}_s)$ is non-convex over $\boldsymbol{r}_i$ and $\boldsymbol{q}_s$. Since the deployment of radar sensors is carried out in an offline manner, the computation complexity is not the main concern of this problem. We thus can use the brutal search over $\boldsymbol{r}_i$ to obtain the near-optimal strategies for both the target and RSN.

2. Pure Strategy: In this case, the target is only allowed to move along one direction at each location. Thus, we define the MSE of the point $\boldsymbol{s}$ as the maximum MSE among those corresponding all the possible observation patterns, i.e.,

$$\gamma_s(\boldsymbol{r}_i) = \max_{1\le j\le J_s} \gamma_s^j, \tag{18}$$

where γ_s^j is given by (13) with only one observation pattern j. Then, the optimal strategy of the target at the point s_s is along moving direction corresponding the j_0-th observation pattern, where $j_0 = \arg\max_{1\le j\le J_s} \gamma_s^j$. Next, the RSN wants to minimize the MSE performance, i.e.,

$$\text{(P2)}\ \min_{\{\boldsymbol{r}_i\}}\max_{\boldsymbol{s}\in\mathcal{S}}\ \gamma_s(\{\boldsymbol{r}_i\}),\ \forall \boldsymbol{s} \in \mathcal{S}. \tag{19}$$

Since $\gamma_s(\{\boldsymbol{r}_i\})$ is non-convex over $\{\boldsymbol{r}_i\}$, Problem (P2) is also non-convex. In general, we can also turn to brutal search over $\{\boldsymbol{r}_i\}$.

5 Numerical Results

In this section, we validate our studies by numerical results. To facilitate the simulations, we divide area $\mathcal{S}$ into some sub-areas, and use one point within each sub-area to represent the whole sub-area. For the purpose of exposition, we adopt the hexagon as one sub-area, and use its center point to denote the corresponding sub-area as $s_{n,m} = \left(x^s_{n,m}, y^s_{n,m}\right)$, $-N \leq n \leq N$ and $-M \leq m \leq M$. It is easy to see that the distance between the point $s_{n,m}$ and each of its six neighbor points is a constant, denoted as τ, where τ is related to the processing time of the RSN to generate one decision. It is easy to check that as τ goes to zero (then N and M go to infinity), the system error introduced by the discrete approximation $\{s_{n,m}\}$ will degrade to zero.

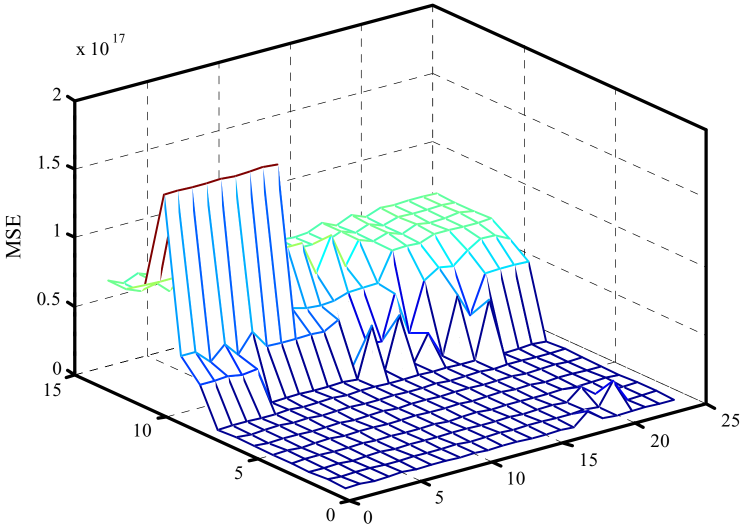


Fig. 3. Maximum MSE performance for different radar sensor location

Furthermore, we consider a rectangular area with the length $50 \cdot \frac{\sqrt{3}}{2}$ meters and the width 30 meters, the distance between any two neighboring points is 1 meter, and there are in total of four radar sensors. For the RSN, the target is filtered out by the MTI when $|\cos\varphi_i| < 0.25$. We set $\sigma_i^2 = 1$ for $i = 1, \cdots, K$. In this subsection, we only consider the pure strategy is utilized at the target. Due to the parameters we adopted in this subsection, it is easy to see that the four radar sensors should be deployed symmetrically in area $\mathcal{S}$, which can shrink the search area of the location of the radar sensors.

In Fig. 3, we plot the maximum MSE over area $\mathcal{S}$ when the first radar sensors located in different points. Due to the symmetry property, we can only consider the case that the first radar sensor is located within the area $[0, 25 \cdot \frac{\sqrt{3}}{2}] \times [0, 15]$. It is easy to observe that the maximum MSE performance can be greatly improved by choosing the optimal location of the radar sensors. For the considered case, the optimal location of the first radar sensor is given by $(\frac{5\sqrt{3}}{2}, 0)$.

In Fig. 4, we plot the MSE performance for each point in area $\mathcal{S}$, and in Fig. 5, we plot the optimal pure strategy of the target at each point. From Fig. 4, it is observed that the points in the center area have a good MSE performance, while the edge areas around the radar sensors are suffering from a much higher estimation error. This is due to the following fact: At these edge points, the target can choose a observation pattern, and be

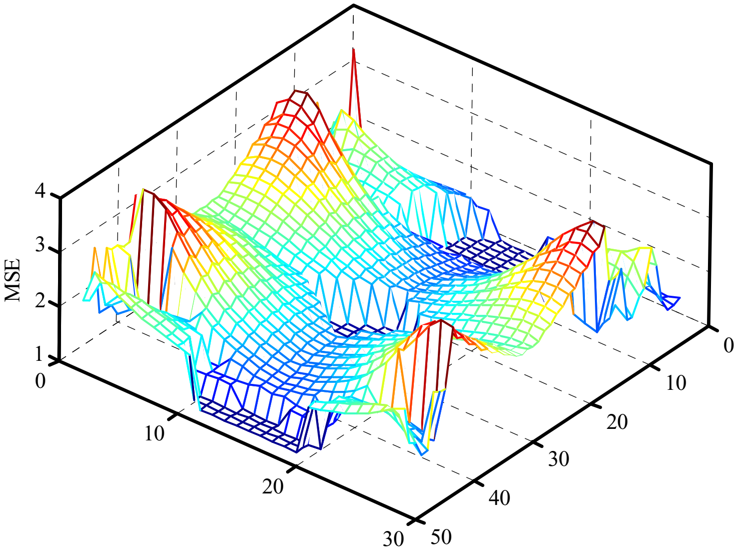


Fig. 4. MSE performance of the localization problem in different points within area $\mathcal{S}$ under the optimal radar sensor deployments

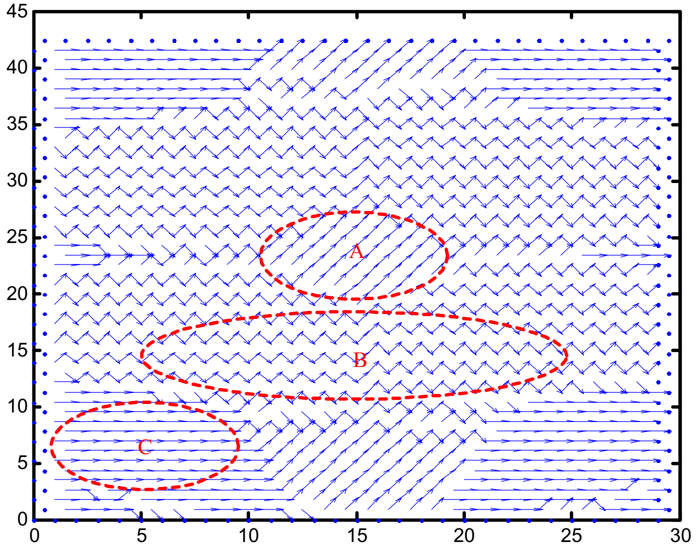


Fig. 5. Optimal pure strategy of the target over area $\mathcal{S}$

seen by certain radar sensors, which contributes less to the determinant of the matrix $\mathcal{J}$ defined in (12). For example, when the target in the points of area C of Fig. 5, it chooses a moving direction by which it can only be observed by the closer radar sensors. On the other hand, the points in the center area of $\mathcal{S}$ own more balanced distances to different radar receivers, and thus the MSE performance of the localization problem is much better than their edge counterparts.

6 Conclusion

In this paper, we investigated the cooperative localization problem in the RSN, by exploiting the information of the observation patterns at the radar receivers to deal with the band-pass effect of the MTI. For the localization in each point of interest, we obtained the CRLB of this problem with general information about the observation patterns. Next, we studied the interaction between the RSN and the target, and formulated the localization as a two-stage Stackelberg game for both the cases that the target adopts the mixed and pure strategies, respectively. We further derived the equilibrium solutions by the principle of backward induction.

References

1. Richards, M.A.: Fundamentals of radar signal processing. McGraw-Hill, New York (2005)
2. Hanle, E.: Survey of bistatic and multistatic radar. IEE Proceeding F, Communications, Radar, and Signal Processing 133(7), 587–595 (1986)
3. Haykin, S.: Cognitive radar networks. In: IEEE Workshop on Sensor Array and Multichannel Processing, Waltham, MA (2006)
4. Sheng, X., Hu, Y.H.: Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks. IEEE Trans. Signal Processing 53(1), 44–53 (2005)
5. Song, X., Willett, P., Zhou, S.: On Fisher information reduction for range-only localization with imperfect detection. IEEE Trans. Aerospace and Elec. Systems 48(4), 3694–3702 (2012)
6. Song, X., Willett, P., Zhou, S.: Target localization with NLOS circularly reflected AoAs. In: Proceeding of ICASSP, Prague, Czech (May 2011)
7. Gong, X., Zhang, J., Cochran, D.: When target motion matters: doppler coverage in radar sensor networks. In: IEEE INFOCOM 2013, Turin, Italy (2013)
8. Huang, C., Chen, X., Zhang, J.: Radar sensor networks: linear cooperative fusion and detection games. Submitted to ICCC 2013 (2013)
9. Mas-Colell, A., Whinston, M.D., Green, J.R.: Microeconomic Theory. Oxford University (1995)
10. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge University Press, Cambridge (2004)

Characterizing the Impact of Non-uniform Deployment of APs on Network Performance under Partially Overlapped Channels

Wei Zhao, Zubair Md. Fadlullah, Hiroki Nishiyama, and Nei Kato

Graduate School of Information Sciences, Tohoku Univeristy, Sendai, Miyagi, Japan
{zhaowei,zubair,bigtree,kato}@it.ecei.tohoku.ac.jp

Abstract. Partially overlapped channels were demonstrated to have the potential of improving the network performance. One example is an increased capacity in a well saturated network. We address the problem of Wi-Fi network planning incorporating partially overlapped channels by more efficiently exploring the spatial reuse to increase the network capacity. We exploit that the interference ranges for separated channels are different, which can be utilized to deploy access points non-uniformly. In this paper, we formulate the problem, show that it can not be solved in polynomial time. Therefore, we propose a greedy optimization algorithm and validate the theoretical results through computer-based simulations.

Keywords: Partially overlapped channel, Wi-Fi network, spatial reuse, channel assignment.

1 Introduction

In recent years, wireless networks have become an increasingly popular field from wireless mesh networks [1], sensor networks [2] to vehicular networks [3] to provide ubiquitous network access to users. IEEE 802.11b/g standards are among the most widely used technology for wireless networks, operating in the ISM 2.4GHz band in which 11 channels are available. The center frequencies are separated by 5 MHz, while each channel occupies a spread of about 30 MHz as shown in Fig. 1.

There are some overlapped frequencies among adjacent channels, also known as the channel interference. This channel interference decreases with the channel separation (CS) which describes the extent of the overlap. With sufficient separation (no less than 5 channels in the IEEE 802.11b standards) no interference will occur. We define channels without frequency overlap as orthogonal channels.

Currently, either one or three orthogonal channels (channel 1, 6 and 11) are employed in Wi-Fi networks. In order to improve network capacity, partially overlapped channels (POCs) were proposed. Recent work shows that a careful design of partially overlapped channels can often lead to significant improvements in spectrum utilization and network performance [5–7].

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 244–254, 2013.

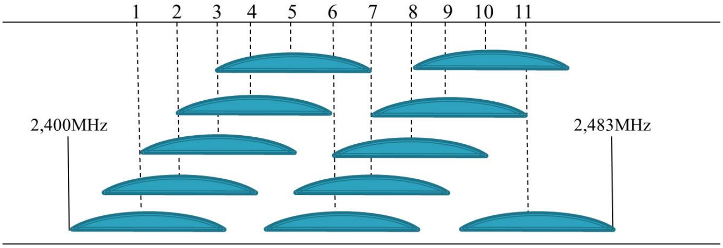


Fig. 1. Frequency spread of various channels in the IEEE 802.11b/g standard[4]

Previous work assumed a uniform[1] [8] or random [4] topology in a network. However, in practical applications, it is infeasible to deploy Access Points(APs) randomly in a WLAN based mesh network and also the placement of APs is restricted by a physical environment. We propose a new scheme combining AP appropriate deployment of APs and channel assignment to improve network capacity.

The contributions of our work are as follows:

1. We consider the practical issue of non-uniformly deploying APs in a one-dimensional topology, for instance, the access network along the subway or metro-rail platform.
2. We propose a greedy algorithm to solve the problem due to computational intractability.
3. Finally, we evaluate the uplink throughput and show via simulations that our scheme outperforms the uniform AP deployment.

The paper is organized as follows. We discuss the related work in Section 2. The problem formulation is described in Section 3. Our proposal combining AP deployment with channel assignment is presented in Section 4. Performance evaluations are given in Section 5. Conclusions and future work are given in Section 6.

2 Related Work

Previous work on POCs ranges from network analysis [5–7, 9] to concrete technologies [4, 10, 8] for the allocation of POCs to APs in practice.

The work in [5], Mishra et al. defined and modelled POCs in wireless environments. The authors measured the amount of partial overlap between two channels from the physical layer and gave the numerical result as interference vector (IV) as shown in table 1. In this paper, we also utilize IV to decide whether two channels interfere.

[1] Here, the definition of uniform deployment is that the distance between Access Points is the same, otherwise, it is non-uniform deployment.

Table 1. An interference Vector conditioned on the channel separation. For instance, for a channel separation of 2 (e.g. channels 3 and 5), the minimal distance for two APs to communicate simultaneously without interfering each other should be at least 190 meters. Also, for a channel separation of 5 or above, no interference is observed even when APs share their location.

Channel Separation	0	1	2	3	4	5	6	7	8	9	10
Distance [meters]	300	280	230	170	70	0	0	0	0	0	0

Mishra et al. also applied the model in the contexts of WLANs and wireless mesh networks with the result that POCs can improve throughput by factors between 1.6 and 2.7. Based on this work, in [9], Feng et al. derived an analytical formulation to calculate the improvement in network capacity compared to utilizing orthogonal channels in networks of string, grid and random topologies. More recently, channel assignment algorithms have been proposed using POCs. A POC-based channel assignment algorithm was proposed in [4] utilizing a new interference model I-Matrix to select channels with less interference. Following this concept, the authors of [10] assigned POCs in wireless mesh networks. By modelling this as a game-theoretic problem, a near-optimum solution could be obtained [8]. However, these methodologies can not be applied in practice in Wi-Fi networks due to the hardness of dynamically detecting the information of radios when a pair of nodes want to communicate with each other. Finally, the Aileron system was proposed in [11], which embeds channel control information in the modulation type so that client and AP need not be tuned to identical channels. This method is feasible to recognise calls of APs and clients.

3 Problem Formulation

In this section, we first discuss the system and interference model before we formulate the problem analytically.

3.1 System Model

With extensive use of smart phones and other wireless devices, such as PDAs and tablet computers, a Wi-Fi network can be rapidly deployed and provide the communication service to cover “the last mile”. Such Wi-Fi hotspots are frequently employed in areas with high user density. APs at single channels can typically provide good communication services to about 20 users. In order to better exploit spatial reuse, APs could utilize POCs as illustrated in Fig. 2.

3.2 Interference Model

We utilize the interference model described in Section 2. We deploy APs non-uniformly since the interference ranges for different channel separations differ greatly. By exploiting this property carefully, we can improve the network capacity. For example, assume interference vector (IV) as

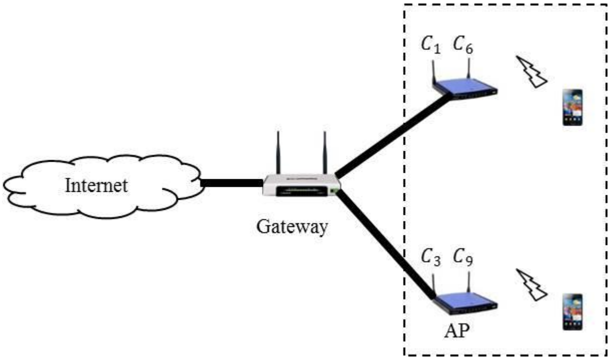


Fig. 2. Wi-Fi network infrastructure utilising POCs. The top AP is assigned channel 1 and 6 while the second is assigned channels 3 and 9.

$$[210\ 190\ 160\ 70\ 10\ 0\ 0\ 0\ 0\ 0\ 0],$$

a communication range of 120 meters and the optimal uniform deployment depicted in Fig. 3(a). When APs are uniformly deployed with a distance of at least 150 meters between neighbours, there are at most 14 channels that can be simultaneously active without interfering with each other as shown in Fig. 3(a). However, with a careful non-uniform deployment, up to 15 channels can be active at a time as shown in Fig. 3(b).

In this paper, we analyze one-dimensional distribution topologies in which all APs are queued in a line. This would, for instance, occur in the access network along a subway platform.

3.3 Formulation of the Problem

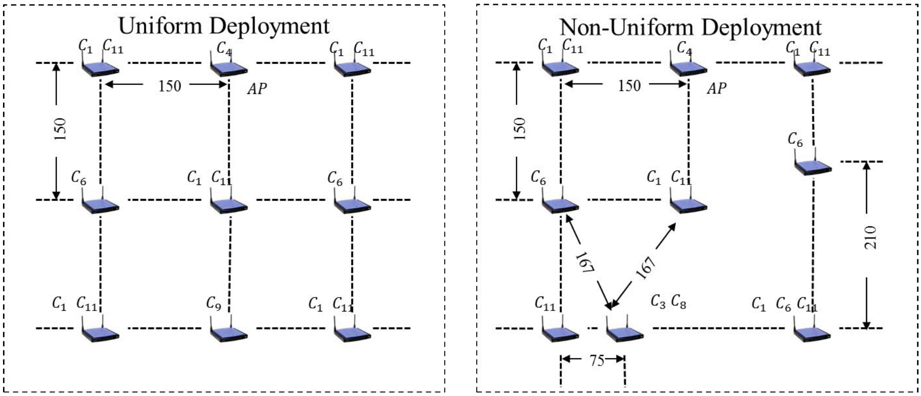
We assume a number of APs of similar capabilities regarding transmission power, the interference ranges and the number N of radios. We assume a high user density so that the uplink traffic from users to APs is delayed. Under such a traffic model, we can consider the uplink throughput without analyzing the interference among users. APs are selected according to their distance.

We model this scenario as an optimization problem with the objective of maximising the overall uplink throughput in a 1-dimensional network. Let $x(m)$ be the distance of the m^{th} AP from the leftmost AP. We define a binary variable c_i^m to indicate the state of AP as

$$c_i^m = \begin{cases} 1 & : \text{ AP } m \text{ transmits on channel } i, \\ 0 & : \text{ otherwise.} \end{cases} \tag{1}$$

Since the uplink transmission from users to APs is congested, we can generalize the objective as maximizing simultaneous transmissions by n APs on M channels (e.g. 11 channels in 802.11b) as

$$\sum_{i \in \{1,\ldots,M\}, m \in \{1,\ldots,n\}} c_i^m \tag{2}$$



(a) Optimal uniform deployment of APs (b) Optimal non-uniform deployment

Fig. 3. Communication channels assigned for two possible deployments of 9 APs and a given interference vector

Some network constraints have to be met to achieve this objective.

First, if channel j is assigned to the m^{th} AP and active, other channels that have some spectrum overlap with channel j can not be assigned to the m^{th} AP. The set of channels overlapped partially or fully with channel i can be donated as $POC(i) = \{max\{1, i-T+1\}, \dots, min\{M, i+T-1\}\}$, where T is the minimum separation for two orthogonal channels. For example, in IEEE 802.11b standard, $T = 5$ and for channel 3, $POC(3) = \{1, 2, 3, 4, 5, 6, 7\}$. Then the orthogonal constraint can be expressed as

$$\sum_{j \in POC(i)} c_j^m \leq 1, \forall i \in \{1, \dots, M\}, \forall m \in \{1, \dots, n\}. \tag{3}$$

In addition, the number of channels on each AP should not exceed the count of radios N equipped in the AP

$$\sum_{i \in \{1,2,\dots,M\}} c_i^m \leq N, \forall m \in \{1, \dots, n\} \tag{4}$$

Because these transmissions are active at the same time, they have to be beyond the interference range of each other as expressed in the constraint 5.

$$\begin{aligned} \left| c_{i+t}^p x_p - c_i^m x_m \right| \geq IR(t) c_{i+t}^p c_i^m, \forall m \in \{1, \dots, n-1\}, \forall p \in \{m+1, \dots, n\}, \\ \forall t \in \{0, \dots, T-1\}, \forall i \in \{1, \dots, M-t\} \end{aligned} \tag{5}$$

Finally, all APs should cover the whole area in order to provide the communication service to all users as

$$0 \leq x_p - x_{p-1} \leq 2R, x_1 \leq R, L - x_n \leq R, \forall p \in \{2, \dots, n\}. \tag{6}$$

Here, R is defined as the communication range and L as the maximum distance between any pair of APs.

In summary, the optimization problem can be formulated as a non-linear programming problem with equation 2 being the objective and equations 3–6 being the constraints. However, although an optimal solution always exists, it is impossible to attempt to optimize the objective by solving the formulation. Given the number of APs, the network area and other network parameters, there must be an optimal solution, which consists of two parts. They are positions for every AP and channels on APs. Even though we know the first part, the optimal positions for APs, the time complexity for solving the maximization problem to obtain the channel assignment is

$$[(n-1)!]^{M-1} + [(n-1)!]^{M-1} +, \ldots, [(n-1)!]^{M-T} = O[(n!)^M] \qquad (7)$$

Therefore, the time complexity for the original problem is greater than $O[(n!)^M]$.

4 Proposed Channel Assignment Technique

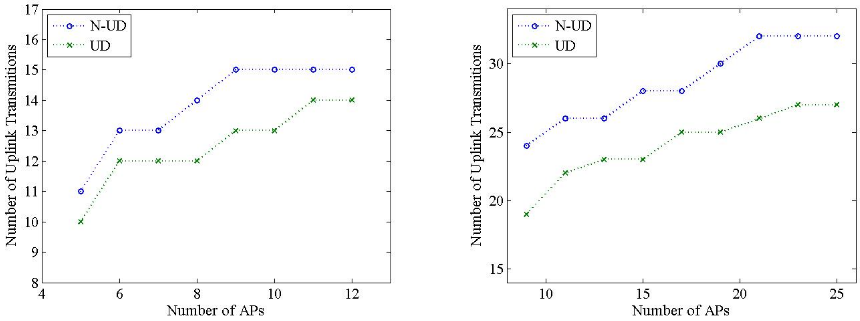
The hardness result in Section 3 provides a compelling reason to investigate heuristic approaches. In particular, we propose a polynomial time greedy algorithm that is able to find a good solution. An optimal solution constitutes locations for every AP and their channel assignment. Since these aspects are not independent, we propose a metric combining them. We define this metric as the channel coverage $CC(x_m) = Num_{ort}(x_m)/(x_m - x_{m-1})$, where x_m is the position candidate for an AP, x_{m-1} is the position of the current AP deployed in the last loop and $Num_{ort}(x_m)$ is the maximized number of orthogonal channels that can be assigned to the AP at the position x_m.

We propose a greedy algorithm to determine AP deployment and channel assignment (algorithm 1). Initially, parameters, such as the interference vector, communication range, dimensions of the placement area are configured (row 2 in algorithm 1). The algorithm then checks whether there are sufficient APs to cover the whole area (from row 3 to row 6). Then, we deploy the first AP and assign channels 1, 6 and 11 (the maximum possible channels). Next, we calculate candidate positions for every channel (from channel 1 to channel 11 in IEEE 802.11b/g standard). Channels with identical candidate locations are grouped (row 13 in algorithm 1). If one group (within the dimensions of the placement area) with the maximum number of channels is found, an AP is deployed there and assigned the channel group $CC(x_{m+1})$. Otherwise, the current number of APs is optimal with respect to the dimensions of the placement area (from row 14 to row 21).

In the case that after the last AP is placed, still not the complete scenario's dimensions are covered, we shift the last APs iteratively to fill the gap on the 1-dimensional area until the complete area is covered (from row 23 to row 45). The first AP is a special case (from row 39 to row 41).

Algorithm 1. Proposed Non-Uniform Deployment Algorithm

```
Initial State
initial Interference Vector (IV), communication range (R), line length L and other perimeters;
if $L/R > (n+1)$ then
    There is no enough AP to cover area.
    return;
end if
deploy the first AP and assign channel;
m=1;
for each $m \in [1, n]$ do
    for channel $j \in [1, NUM\_CHANNEL]$ do
        calculate the candidate position for channel $j$;
    end for
    divide candidates into groups, each group includes channels that are assigned at the same position;
    find the maximized $CC(x_{m+1})$;
    if position $x_{m+1}$ exist then
        deploy AP at $x_{m+1}$;
        assign channel to the $(m+1)^{th}$ AP;
    else
        $m$ is the maximized number of AP that be able to be deployed in the area;
        return;
    end if
end for
$m = n$;
if $x_m + R \leq L$ then
    while $m \geq 2$ do
        if $m == n$ then
            $x_m = last\_position - R$;
        else
            $x_m = last\_position - 2R$;
        end if
        assin channel $1, 6, 11$ to the AP at $x_m$;
        if $x_m - x_{m-1} \leq 2R$ then
            return;
        else
            $last\_position = x_m$;
            $m = m - 1$;
        end if
    end while
    $x_1 = x_2 - 2R$;
    assign channel 1, 6, 11 to the AP (at $x_1$);
    return;
else
    the $n^{th}$ AP covers the area;
    return;
end if
```



(a) Comparison of UD and N-UD performance with a 1000m placement area

(b) Comparison of UD and N-UD performance with a 2000m placement area

Fig. 4. Simulation results for 1000m and 2000m placement areas

5 Performance Evaluation

To demonstrate how non-uniform deployment for APs can be used to improve network capacity, we explore maximized simultaneous uplink transmissions as the metrics to evaluate the performance of our model and our proposed algorithm. In particular, we compare the algorithm with the uniform deployment of APs for varying number of available nodes and varying dimensions of the placement area. For the uniform deployment, we assume that the first AP is placed at one end of the deployment area and the last AP at the other end. In order to study the uniform deployment (UD), we set the distances among all APs in our model as identical. This problem can then be solved by "brute-force"in Lingo [12] when the number of nodes in the model is less than 30.

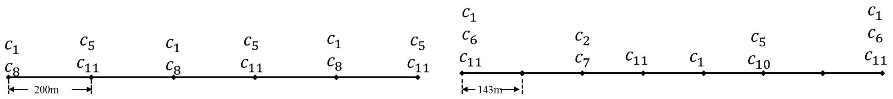
Apart from UD, we consider non-uniform deployment (N-UD) and the traditional channel assignment (UD-OC) employing orthogonal channels (namely channel 1, 6 and 11 in IEEE 802.11b/g standards).

Fig. 4(a) plots the maximized number of uplink transmissions achieved by N-UD and UD when the length of the placement area is 1000 meters, the valid communication range is 150 meters and the interference vector is

$$[300 \quad 280 \quad 230 \quad 170 \quad 70 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0].$$

In order to cover the whole line segment, we have to deploy at least 5 APs. As the figure shows, when the node number is relatively small (5 to 7), there is, in compliance with the paper [11], just a minor improvement for N-UD over UD. With higher AP density (8-9), the interference greatly increases, with negative impact on simultaneous uplink transmissions while in contrast N-UD can make full use of the spatial freedom and performs much better than UD.

Also, as more APs are added into the environment, there is no improvement for N-UD because the optimum number of APs that can be deployed in the



(a) Six APs placed in uniform distance and their available channels

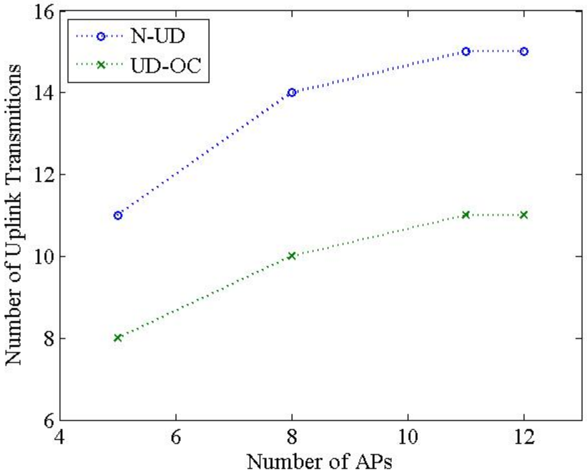
(b) Eight APs placed in uniform distance and their available channels

Fig. 5. Some performance in uniform deployment for different numbers of APs when the line length is 1000 meters

area is 9. On the other hand, with UD it often happens that there will be no improvement by increasing the number of APs as the example shown in Fig. 5.

In Fig. 5(a), there could be 12 uplink transmissions in total. However, by adding one or two further APs would not increase the performance as shown in Fig. 5(b) since the distance among APs is fixed and therefore does not provide any optimization potential. Comparable results here achieved for a 2000 meter placement area as shown in Fig. 4(b), in which N-UD also outperforms UD and shows the similar performance with Fig.4(a).

In addition, we compared N-UD and UD-OC in Fig. 6 for 1000 meter placement area. Note that there are fewer channels in UD-OC than POCs, it shows worse performance in UD-OC than others.



(a) 1000 meters placement area

Fig. 6. Comparison result between non-uniform deployment and uniform deployment for 1000 meter placement areas

6 Conclusion

In this paper, we considered the non-uniform deployment for APs employing POCs in Wi-Fi networks to improve network capacity. We provided an analytical model, derived its non-polynomial time complexity and proposed an alternative greedy polynomial heuristic for AP deployment in this setting. Our conducted simulation results reveal that the scheme gains capacity improvement over uniform deployment, e.g., along the subway platform. Future work will extend the work to two-dimension areas.

Acknowledgement. This work was conducted under the national project, Research and development of Movable ICT-Units for emergency transportation into disaster-affected areas and multi-unit connection, supported by the Ministry of Internal Affairs and Communications (MIC), Japan. And We thank Stephan Sigg for his helpful comments.

References

1. Akyildiz, I.F., Wang, X.: A survey on wireless mesh networks. IEEE Communications Magazine 43(9), S23–S30 (September)
2. Yick, J., Mukherjee, B., Ghosal, D.: Wireless sensor network survey. Computer Networks 52(12), 2292–2330 (2008)
3. Biswas, S., Tatchikou, R., Dion, F.: Vehicle-to-vehicle wireless communication protocols for enhancing highway traffic safety. IEEE Communications Magazine 44(1), 74–82 (January)
4. Hoque, M.A., Hong, X., Afroz, F.: Multiple radio channel assignement utilizing partially overlapped channels. In: IEEE Global Telecommunications Conference, GLOBECOM 2009., pp. 1–7 (December 4, 2009)
5. Mishra, A., Shrivastava, V., Banerjee, S., Arbaugh, W.: Partially overlapped channels not considered harmful. In: Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS/Performance 2006, ACM, New York (2006)
6. Mishra, A., Rozner, E., Banerjee, S., Arbaugh, W.: Exploiting partially overlapping channels in wireless networks: turning a peril into an advantage. In: Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement, IMC 2005, pp. 29–29. USENIX Association, Berkeley (2005)
7. Feng, Z., Yang, Y.: How much improvement can we get from partially overlapped channels? In: IEEE Wireless Communications and Networking Conference, WCNC 2008, March 31-April 3, pp. 2957–2962 (2008)
8. Duarte, P.B.F., Fadlullah, Z.M., Vasilakos, A.V., Kato, N.: On the partially overlapped channel assignment on wireless mesh network backbone: A game theoretic approach. IEEE Journal on Selected Areas in Communications 30(1), 119–127 (2012)
9. Feng, Z., Yang, Y.: Characterizing the impact of partially overlapped channel on the performance of wireless networks. In: IEEE Global Telecommunications Conference, GLOBECOM 2008, pp. 1–6 (December 42008)

10. Duarte, P.B.F., Fadlullah, Z.M., Hashimoto, K., Kato, N.: Partially overlapped channel assignment on wireless mesh network backbone. In: IEEE Global Telecommunications Conference (GLOBECOM 2010), pp. 1–5 (December 2010)
11. Liu, H., Yu, H., Liu, X., Chuah, C.-N., Mohapatra, P.: Scheduling multiple partially overlapped channels in wireless mesh networks. In: IEEE International Conference on Communications, ICC 2007, pp. 3817–3822 (June 2007)
12. Lingo, `http://www.lindo.com/`

Patient's Motion Recognition Based on SOM-Decision Tree

Wei Yu, Hongli Yan, Junqi Guo*, and Rongfang Bie

College of Information Science and Technology, Beijing Normal University, Beijing, China
weiyu@cupes.edu.cn, yanhongli418@163.com,
{guojunqi,rfbie}@bnu.edu.cn

Abstract. Patient's motion recognition is quite popular in the area of healthcare and medical service nowadays. By analyzing the data from variant sensors within the network, we can estimate the activities a person does. The analyzing job is usually done by a classifier which can classify each motion into one category with similar movements. Self-Organizing Map (SOM) is a kind of algorithm that can be used to arrange data into different categories without any guidance. Decision tree is a mature tool for classification. In this paper, we propose a new kind of classification method with data from BAN called SOM-Decision Tree. Firstly, we use SOM on each of the sensor nodes to categorize motions into different classes, so that motions in different classes can be distinguished by this sensor. Secondly, a decision tree is constructed to discriminate each kind of movements from other motions. Finally, any action of the same patient can be recognized by query through the decision tree. According to our experiment, this algorithm is feasible and quite efficient.

Keywords: Motion recognition, SOM, Self-organizing Map, Decision Tree, classification, Mobile Health.

1 Introduction

Patient's state monitoring is a common task for medical staffs, for example, high-risk infants need to be observed day and night[1], and patients with hemiplegia should be paid attention to due to their reduced mobility. However, it becomes a heavy burden when the number of patients increased. Therefore, we need to employ information techniques to assist the detection of human status, such as the application of moving cameras in human motion detection[2]. As the development of the techniques of the Internet of things, many wearable (including implantable) wireless sensors and equipments are used in monitoring human states[3,4]. With the assistance of these technologies, patients in hospital as well as those stay at home can be monitored equally. In areas of health care and fitness, many systems have been designed to recognize and evaluate human status with the application of wireless sensors[5,6]. Among these studies, Body Area Network(BAN) is widely used to acquire human motion data. The concept of BAN was first proposed by T. G.

* Corresponding author.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 255–268, 2013.

Zimmerman and defined in the wireless World Research Forum's Book of Visions as "a collection of (inter) communicating devices which are worn on the body, providing an integrated set of personalized services to the user"[7]. As the development of wireless communication technology, BAN is extended to be wireless body sensor network(WBAN)[8]. In some study, it is also defined as Body Sensor Network(BSN)[5]. It can be used to detect human motion status, such as motion recognition[9], intervention of patient activation[10].

In order to identify different human movements, classification algorithms are commonly used, such as HMM, Bayesian classifier, SVM, etc. for example, [11,12] used decision tree to detect user states, [13] compares a reference majority voting and a naive Bayesian fusion scheme with HMM algorithm in classification for activity recognition from on-body sensors, [14] used a Bayesian classifier with multivariate Gaussians to model patient's activity, [15] compared supervised and unsupervised physical activities using a hybrid classifier combining a tree structure containing a priori knowledge and artificial neural networks. Self-organizing Map(SOM) is a sheet-like artificial neural network, that cells of which become specifically tuned to various input signal patterns or classes of patterns through an unsupervised learning process[16].It can be used as a tool in pattern recognition[17-19], clustering[20-22], classification[23-25] as well as data visualization[26]. As a cluster algorithm, SOM has been widely used in many practical areas, such as machinery health monitoring[27,28], which has proved the feasibility, practicability and priority of this algorithm as a clustering algorithm. In mobile health area, there were a few attempts of using SOM, for example, detecting the turning points of human activity based on wearable sensor array data[29], [30] used accelerometer to recognize activities of children in kindergarten, and evaluated the performance using some orthodox classifiers together with SOM, [31] introduced a visualization method for activity information sharing system using Self-Organizing Map and the activity information sharing system "ALKAN2". If we label the data in the map, SOM can also be used as a classifier. For example, in [32,33], Self-Organizing Map is used as a tool to classify a person's motion into different categories, which was the base of motion classifier that used to identify the action of fall. In this paper, we propose an algorithm that uses SOM as a clustering tool on each nodes of WBAN, which sorts different motions into various categories. As SOM is applied on each individual nodes of WBAN, we can perform the algorithm concurrently if condition permits, which increases the efficiency that matters a lot as for online algorithm. Then a decision tree[34] is constructed to perform the division of all the motions based on the results of the clustering on accelerometers within WBAN. The order of the nodes within the decision tree is chosen based on greedy strategy. According to [34], this distinction is proved to be NP-complete. Thus, we can use this method to detect patients' body motion so as to monitor their status.

Contents of the paper are as follows. In section 2 we give an overview of the basic concepts of algorithms that involved in our study. Section 3 describes the proposed algorithm in detail with its novel improvements and analysis of its application in medical care area. It is followed by several qualitative and quantitative experiments that aim to prove the feasibility, high efficiency and priority of our algorithm strategy in section 4. Then the application of this strategy in human motion recognition within medical background is simply discussed in section 5. Section 6 concludes the whole study of this paper and prospects the future work.

2 Architecture of the Motion Recognition Scheme

Movement analysis is a popular topic with a wide range of applications in health care related areas, and accelerometer is the most popular sensor applied in motion detection applications[30,35]. In our study, we focus on recognition of body motion with an array of accelerometers that form a WBAN. We design a motion recognition scheme for medical purpose which can identify various motion types of patients being observed. This scheme concentrates on discrimination of different motion types within a short period of time so that it can be applied as an online solution for medical health monitoring.

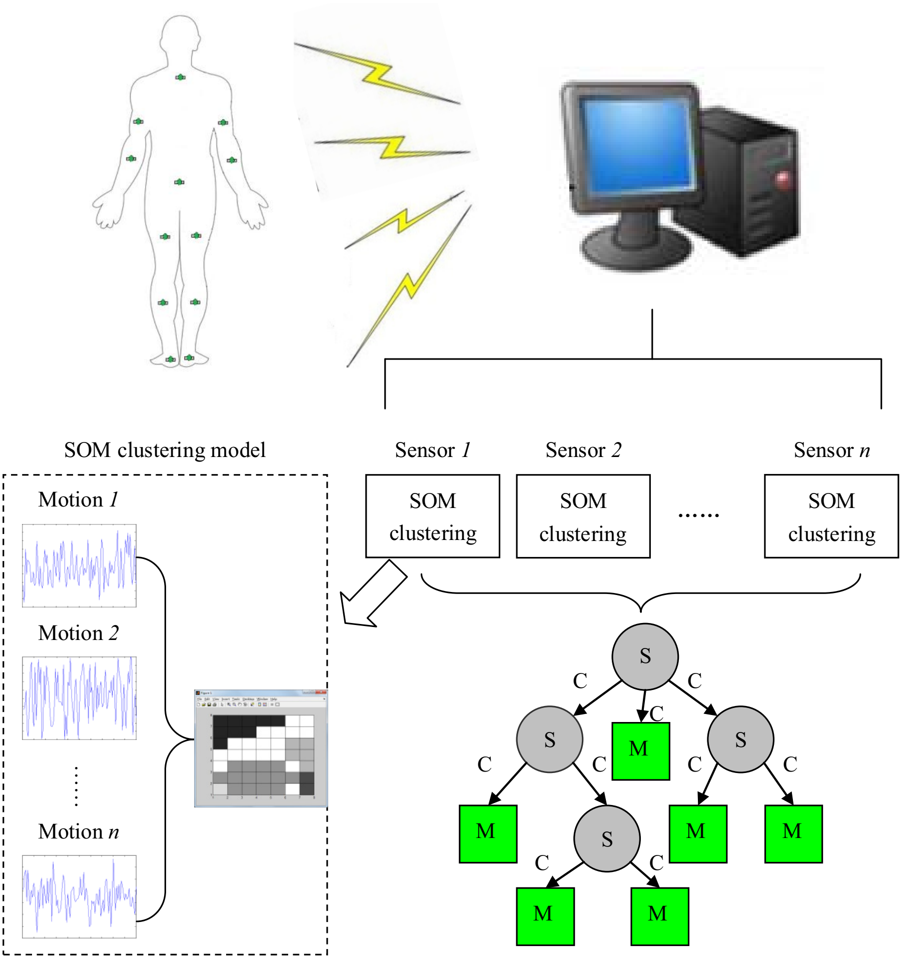
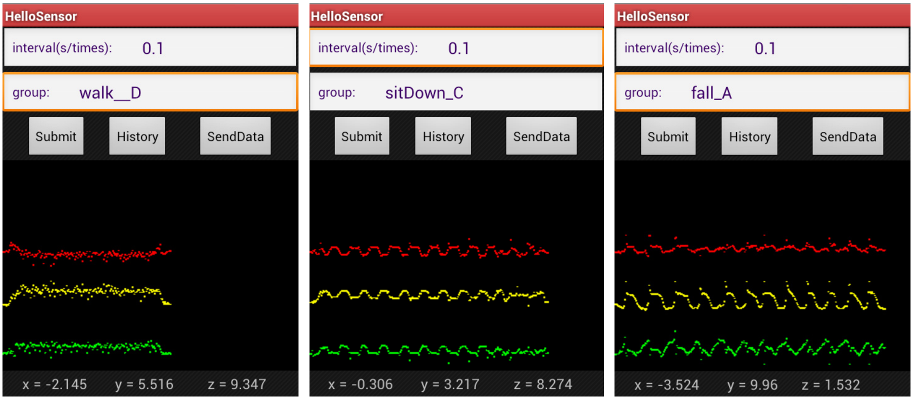


Fig. 1. Structure chart of the scheme

Our motion recognition scheme intend to be applied as a solution that employs several acceleration sensors which possess three dimensions *x*, *y* and *z*-axes along three directions as nodes of the body area network. So we use several smart phones with accelerometers to simulate motion data of sensors within the BAN refer to [32]. Data from the nodes are processed separately so as to distinguish different motion effects on the same body part. Inspired by the former study[29], we decided to analyze each nodes independently, and then choose a few with high distinguish degree with greedy policy. The structure chart of the scheme is shown in Figure 1.

The whole process of our strategy can be summarized as follow.

Data Simulation: According to different aims, different ranges of sensors should be chosen. We place 4 smart phones on different part of human body, like arm, leg, hand as 4 nodes for recognition of 8 normal actions (jogging, walking, walking down the stairs, climbing the stairs, fall, stand-sit-stand, squatting down, lying down). As the limitation of sensors in smart phones, a few simulation is done before the preprocessing. The motion data of walking, stand-sit-stand and fall acquired from the smart phone is shown in Figure 2.



HelloSensor
interval(s/times): 0.1
group: walk__D
x = -2.145 y = 5.516 z = 9.347

HelloSensor
interval(s/times): 0.1
group: sitDown_C
x = -0.306 y = 3.217 z = 8.274

HelloSensor
interval(s/times): 0.1
group: fall_A
x = -3.524 y = 9.96 z = 1.532

Fig. 2. Motion data of walking, sitting and fall acquired from smart phone

Data Preprocessing: Data from various sensors are preprocessed for further clustering and classification. According to [32], we transform the three dimensional data array $\left(a_x, a_y, a_z\right)$ into $\left(a_i, t_i\right)$ array which indicates the registration of the sensor at time stamp t_i. Here, a_i can be calculated as below:

$$a = \sqrt{a_x^2 + a_y^2 + a_z^2} \tag{1}$$

In order to extract the detailed feature, we pick 30 time stamps within 3 seconds to form a data vector that represent a certain motion process in given time interval.

Data Clustering on Each Node: By dividing different motions into various groups based on their diversities on each single node, we can initially distinguish a few movements on one node, while some motions appear to be similar to other motions on certain nodes. Here in this case, we use SOM as the clustering algorithm to categorize motions on each node. As the clustering processes within each node are independent to each other, so we can perform them concurrently of the processing element of the server permits.

Motion Classification: Results of the clustering are input for further discrimination. The aim of the classification is to identify the motions within the same cluster separately from other nodes, so that all motions can be recognized. Here, we use decision tree to do the classification, and the nodes are selected using greedy strategy which has been proved to be practical in [29].

Motion Identification: After the construction of the decision tree, the classifier is ready to work. New motions can be recognized by our scheme.

3 Data Analysis Strategy and SOM-Decision Tree

Our strategy of identifying different motions by arranging them into various defined categories is to distinguish different motions roughly on each sensors of the BAN using SOM and synthesize the clustering results by discriminates motions in the same clusters using a decision tree. We describe the whole algorithm as SOM-Decision Tree, which can be used as a universal scheme for motion reorganization in BAN. This algorithm is mainly composed of two parts: clustering motions into various categories on each sensor and classifying all motions by constructing a decision tree that can distinguish motions within the same category. In the first part, as the operations of clustering on each sensor are independent to each other, we can apply concurrent computation in this process if the hardware condition is available, so that it would be more efficient with short consuming time. In the second part, we can select the most effective essential sensor array that differentiates all motions, especially the motions within the same cluster. This section is going to introduce this algorithm in details within the human motion reorganization background discussed in chapter 2.

3.1 Clustering with SOM on Each Sensor

The first step of our solution is to apply SOM on each sensor concurrently and cluster motions into different categories. SOM is the abbreviation of Self-Organizing Map, which is an unsupervised learning paradigm in artificial neural network[12]. Its main idea is to find the Best Matching Unit (BMU) and update the neural network, which is done by mapping high dimensional input vectors to analogical neurons, and modifying the weights of the best matching neuron as well as its neighbors according to their distance to the hit neuron center. After certain amount of iterations of similar training, the output neural network is ready to distinguish different kinds of input vectors, and cluster them into different categories. Here, the similarity between input data and neurons are scaled with space ranging methods, such as Euclidean distance, which is the most common way being used.

The process of the algorithm is consisted of two major steps: competition stage and cooperation stage. The first stage is also called self-organizing step, or ordering step. In this stage, the random output neurons compete with each other to become the best matching unit, which can also be treated as an ordering process for output data to be organized as the distribution of the input data. The nearest neuron of the input vector is chosen to be the BMU in each iteration, which can be understood as the neuron wins the competition with other neurons in matching with the input vector. After that, the weights of some neurons are modified in the second stage, which is also called the convergence stage. In this step, the winner has the right to modify its weight so as to be closer to the future cluster center, and its neighbors can also change their weights to some extents according to their distance to the BMU. This job is done by calculating the weight adjustment function selected by users themselves.

Concretely, the procedure of SOM algorithm can be described as follow:

1. Assigning small random numbers to the weight vectors of the output layer and performing normalization. After that, the output vector is obtained as $w_j = [w_{j1}, w_{j2}, \ldots, w_{jn}]^T, j = 1,2,\ldots,l$, l is the number of output neurons. The input pattern after normalization can be indicated as $x = [x_1, x_2, \ldots, x_n]^T$, n is the number of sensors.
2. Competition stage: Finding the Best Matching Unit (BMU) c using Euclidean distance method.

$$c = \arg \min_j \{\|x - w_j\|\} \tag{2}$$

Here, the Euclidean distance is calculated as:

$$\|x - w_j\| = sqrt\left[\sum_{i=1}^{n}(x_i - w_{ji})^2\right] \tag{3}$$

3. Cooperation stage: Modify the weight value of the BMU and its neighbors. The weight adjustment function is showed as follow:

$$w_j(t+1) = w_j(t) + \alpha(t)h_{cj}(t)(x - w_j(t)) \tag{4}$$

Here, $\alpha(t)$ is the learning rate that can be defined as below (α_0 is the initial value of learning rate, and τ_1 is the exponential decay function of time):

$$\alpha(t) = \alpha_0 \exp\left(-\frac{t}{\tau_1}\right) \tag{5}$$

$h_{cj}(t)$ is the neighborhood function that reveals the distance between BMU and current neuron. We decide to use the function shown in formula 6 which has been commonly used before and appears to be effective.

$$h_{cj}(t) = \exp\left(-\frac{d_{c,j}^2}{2\sigma(t)^2}\right) \tag{6}$$

Here, the distance $d_{c,j}$ is also calculated by Euclidean distance, and the neighborhood radius can be calculated using formula 7 that σ_0 is the initial value of the neighborhood radius, and τ_2 is another exponential decay function of time.

$$\sigma(t) = \sigma_0 \exp\left(-\frac{t}{\tau_2}\right) \tag{7}$$

It is easy to notice that both of the learning rate function and neighborhood function are annealing function.

4. If stable feature mapping is formed, then our learning is finished, otherwise, we should return to step 2 after setting $t = t + 1$.

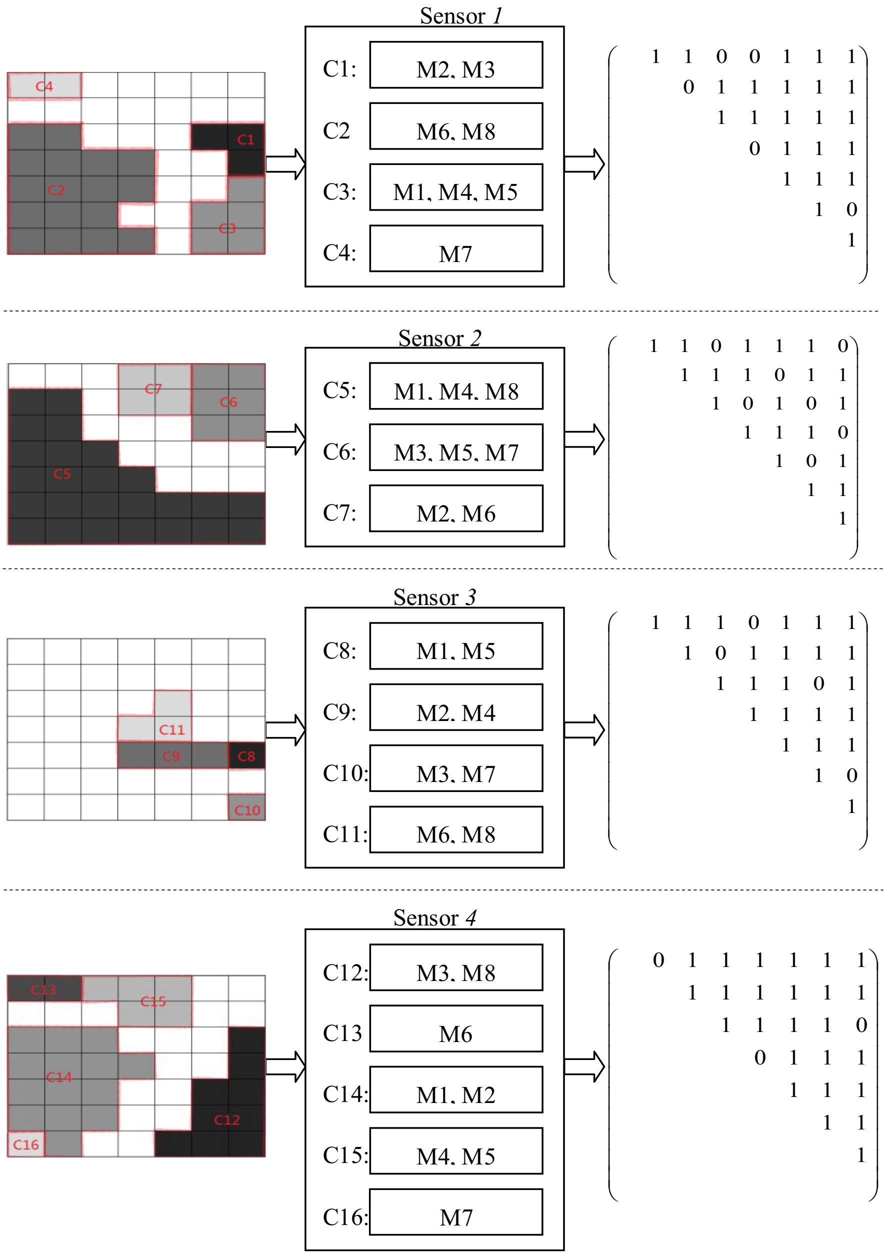


Fig. 3. Results of the SOM clustering and discrimination matrixes

As mentioned before, we select 4 nodes to distinguish 8 motions. Data from different sensors are captured and calculated using formula 1. After standardization, they will be the input vectors of the SOM clustering algorithm. The result of this step is showed in Figure 3.Here, the cells filled with the same color are considered to be in the same category. In order to check whether the current sensors are adequate to discriminate all motions, as well as determine the order of sensors in the decision tree, we construct a bunch of right upper triangular matrixes to indicate the discrimination situation of each sensor. These matrixes are easily built by finding out all the motion pairs within the same cluster and setting the corresponding elements of the matrix to 0, and the rest of the matrix elements are simply assigned with certain values according to their locations following the rules described in the previous section. The discrimination matrixes are also showed in figure 3.

3.2 Construction of the Decision Tree

The second step of our strategy is to construct a decision tree based on the clustering results of each node as well as the discrimination relations between motions represented by respective upper triangular adjacency matrix called discrimination matrix of each sensor. The possibility of the discrimination between all motions we discussed is assumed to be without doubt in [29]. So they focus on finding the complete ordering. However, here in our study, we decide to choose the essential sensors from all sensors in our BAN, and prove the distinguish ability using the complete discrimination criterion defined later in this section. If the existing sensors are proved to be adequate to distinguish all the input motions, then we can start our next step of decision tree construction. Otherwise, we might need to remind user to add more sensors to differentiate current motions. In this way, we don't need to configure too many sensors at first, as sensor number can be supervised by a few pre-experiments.

In [30], the discrimination relations are described by calculating the distinguished pairs of actions (LDS_i) on each node. However, we find it costs too much to calculate all the LDS_i, and counting the pairs of actions that can not be differentiated is easier with the same effect as the total number of relations between motions are equal in all sensors. So we choose to search indistinguishable pairs of actions by looking into the categories, as motions in the same category can not be distinguished. Other than these pairs, all the motions pairs can be treated as discriminated. Therefore, we propose a kind of right upper triangular matrix to describe the discriminated relation between motions on a certain sensor, and we called it the discrimination matrix of the sensor. Here, $w_v(i,j)$ indicates the discrimination situation between motion i and j on sensor v, that 1 indicates distinguishable, and 0 indicates indistinguishable. In this matrix, we only need to care about the right upper triangular part with $i > j$, as the matrix is symmetrical. Go through all the clusters within sensor v, we can set $w_v(i,j) = 0$, if motion i and j are within the same cluster. After filling out all the 0s, we can give 1 to the rest elements in the upper triangular part of the matrix. The discrimination matrixes of all sensors are organized to form a three-dimensional matrix W with $W(v,i,j)$ indicates the discriminated relation between i^{th} and j^{th} motions on v^{th} sensor. The establishment of matrix W is equal to the calculation of GDS and all the *LDS*. In order to ensure the number of our sensors is adequate enough to discriminate all the motions, a calculation should be done first, and our proof of the distinguish ability of our BSN is performed at the same time. This verification is done by consulting of the complete discrimination criterion below:

Complete Discrimination Criterion: Let $W' = w_1 + w_2 + \ldots + w_r$ (r represents the number of sensors), if there exists any 0 in the upper right triangular part of W', then this sensor array is not adequate to discriminate all the motions, otherwise, the sensor array is enough to distinguish the provided motions.

After the construction of W and the proof of the distinguish ability of the sensor network, a decision tree is constructed based on matrix W . Decision tree is a tree type data structure that contains two kinds of nodes: sensor node and motion node. In a decision tree, a sensor node can be either a root node or an internal node, and motion node must be a leaf node. Sensor nodes are parent nodes that possess links to child nodes, and the number of its child nodes is equal to the number of its categories. In order to unify the structure of all sensor nodes, we set $n-1$ links in its structure, because the maximum number of clusters on a sensor is n . That is to say, there might be vain links in a sensor node structure, and the number of valuable links is determined by the category number. Motion node is a leaf node without any child nodes. They are the end of the decision making process which shows the judgment of a given unknown action.

The core process of the construction of decision tree is the generation of branch nodes, especially the production of sensor nodes, which is done by choosing the most distinguishable sensor node for each branch to discriminate the motions within the corresponding category. Here, we can check the sensor matrix value of the involving motions of its parent node to estimate whether a sensor is distinguishable on certain branch. If the discrimination situation between motions in parent node shows any variation from ancient nodes, which means there exists transmission from 0 to 1, then the sensor is distinguishable, otherwise, we should choose the next node from matrix O instead of current node, as it can not distinguish motions within this category.

We use a greedy strategy to construct the decision tree, the details of which are showed as below:

```
Inputs: A, C, S     //A: motion set; C: category set; S: sensor set
Output: T     //T: the decision tree
```

1. For each sensor s_k , build the discrimination matrix w_v as the sub-matrix of matrix W , which meets the following conditions (Here, c_x indicates cluster on sensor v):

$$w_v(i,j) = \begin{cases} 1 & (i > j \wedge (i \in c_x \rightarrow j \notin c_x)) \\ 0 & (i > j \wedge (i \in c_x \rightarrow j \in c_x)) \end{cases} \tag{8}$$

The value of the upper triangular part of the adjacency matrix on one sensor is configured by finding all couples of motions that belong to the same category and set to 0, and the rest can be set to 1.

```
2.  Calculate the number of zeros in each sensor's adjacency matrix, and arrange the
    sensors' serial numbers in accordance with the number of zeros in reverse order in
    matrix $O$. The one with least zeros ranks first and the one with most zeros comes
    in last.
3.  A three-dimensional matrix $W$ is built using the following formula:
        $W(v,i,j)=w_v$                                                          (9)
4.  If the sensors satisfy the complete discrimination condition
5.  then go to 7
6.  else return and add more sensors
7.  Build a stack $E$ for sensor nodes that needed to be further developed.
8.  $k=1$
9.  Build a node $S_k$ using the information from $w_{O(k)}$, and push it into the stack $E$.
10. $T=S$
11. While ( $E \neq \phi$ )
12.     Pop an element $S_k$ out of $E$
13.     Find categories of sensor $s_k$ that involving motions in current node $S_k$
14.         for each of these category of sensor $s_k$
15.             If this category contains only one common motion with $S_k$
16.             Then build a motion node and link it to current node as its child node
17.             Else
18.              $y=k+1$
19.              While node in $O(y)$ is indistinguishable
20.                  $y=y+1$
21.              end while
22.            Build a sensor node $S_{O(y)}$, and link it to current node as its child node
23.            Push $S_{O(y)}$ into the stack $E$
24.            end if
25.         end for
26. end while
```

We use different shapes to represent them separately so as to be distinguished more clearly in the tree. Using the output of the first step as the input of our algorithm, the decision tree of the example we are talking about is showed as below (in Figure 4).

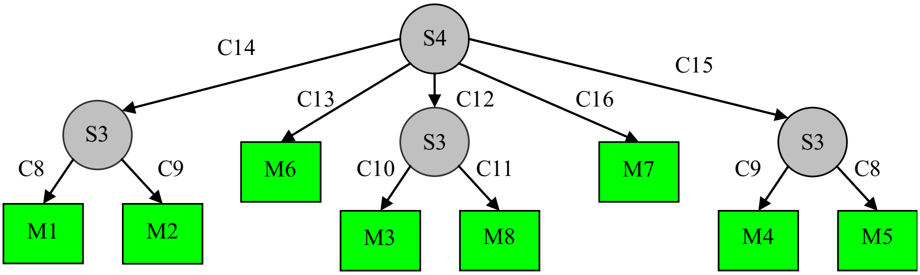


Fig. 4. The decision tree of our example

4 Simulation Analyses and Result Discussion

In order to test the feasibility and priority of the algorithm we proposed, a few simulation and experiments are done. We will compare our algorithm with two relative algorithms which have already been widely used: pure SOM and K-means-Decision tree. In our experiment, we simulate 4 sensors to distinguish ordinary motions. We use these sensors to discriminate 8 kinds of daily actions and compare the accuracy of the three methods. The result is showed in Figure 5. We find that the accuracy of SOM-Decision Tree is the highest, as SOM can discover the accurate similar motion categories that implied in each sensor without any priori knowledge. Thus, the clusters we obtained can match the actual situation more closely. Although the result of K-means Decision Tree is also quite good, it is seriously affected by the selection of K.

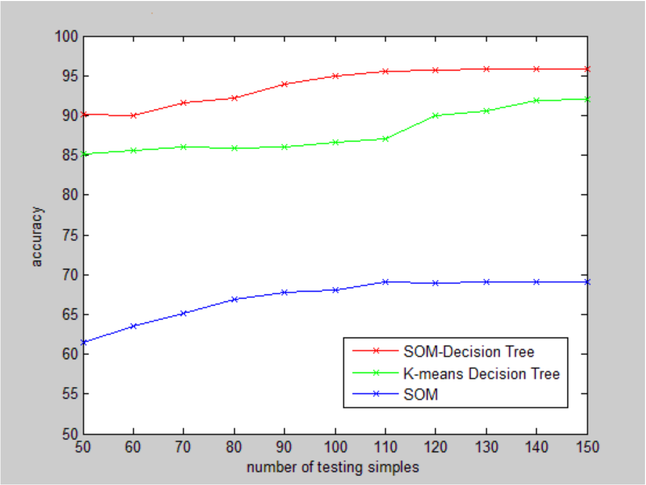


Fig. 5. The comparison of the accuracy of three algorithms

The shown result display the fortunate circumstance of finding the right K, however, you might have chosen the wrong one and the performance would be shameful. In this situation, priori knowledge seems to be extremely important. However, we can not ensure to acquire enough information to choose the right number of clusters, so the performance will not be able to compete with our algorithm. Performance of SOM seems to be barely satisfied for its weak distinguish ability.

According to our experiment, 4 sensors are sufficient for recognition of 8 motions, so the workload is within the tolerance range. As the clustering can be done concurrently, the time consuming is better than pure SOM.

5 Conclusions and Future Work

In this paper, we propose an algorithm SOM-Decision Tree that combines SOM and Decision Tree to classify different human motions and intend to apply it in mobile health area. The feasibility and priority of this algorithm is easily proved by our experiment. In medical field, patient motion recognition is quite common, and it can be used for its short consuming time and high correctness.

In fact, this algorithm can be applied not only in patient motion identification, but also in many motion or gesture recognition fields. So our future work is to spread its application and build a system that can be applied in real scene.

Acknowledgement. This research is supported by National Natural Science Foundation of China (61171014) and the Fundamental Research Funds for the Central Universities.

References

1. Hayes, G.R., Patterson, D.J., Singh, M., Gravem, D., Rich, J., Cooper, D.: Supporting the transition from hospital to home for premature infants using integrated mobile computing and sensor support. Personal and Ubiquitous Computing, doi:10.1007/s00779-011-0402-4
2. Cutler, R., Davis, L.: Robust Real-Time Periodic Motion Detection, Analysis, and Applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 781–796 (2000)
3. Ståhl, O., Gambäck, B., Turunen, M., Hakulinen, J.: A Mobile Health and Fitness Companion Demonstrator. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session, pp. 65–68 (2009)
4. Xu, F., Qin, Z., Tan, C.C., Wang, B., Li, Q.: IMDGuard: Securing Implantable Medical Devices with the External Wearable Guardian. In: IEEE INFOCOM, pp. 1862–1870 (2011)
5. Shahriyar, R., Bari, M.F., Kundu, G., Ahamed, S.I., Akbar, M.M.: Intelligent Mobile Health Monitoring System (IMHMS). International Journal of Control and Automation 2(3), 13–28 (2009)
6. Bourouis, A., Feham, M., Bouchachia, A.: Ubiquitous Mobile Health Monitoring System for Elderly (UMHMSE). International Journal of Computer Science & Information Technology 3(3), 74–82 (2011)

7. Jones, V., van Halteren, A., Widya, I., Dokovsky, N., Koprinkov, G., Bults, R., Konstantas, D., Herzog, R.: Mobihealth: Mobile Health Services Based on Body Area Networks. In: M-Health Emerging Mobile Health Systems, pp. 219–236 (2006)
8. Ullah, S., Higgins, H., Braem, B., Latre, B., Blondia, C., Moerman, I., Saleem, S., Rahman, Z., Kwak, K.S.: A Comprehensive Survey of Wireless Body Area Networks. Journal of Medical Systems 36(3), 1065–1094 (2012)
9. Wu, C., Tseng, Y.: Data Compression by Temporal and Spatial Correlations in a Body-Area Sensor Network: A Case Study in Pilates Motion Recognition. IEEE Transactions on Mobile Computing 10(10), 1459–1472 (2011)
10. Solomon, M., Wagner, S.L., Goes, J.: Effects of a Web-Based Intervention for Adults With Chronic Conditions on Patient Activation: Online Randomized Controlled Trial. Journal of Medical Internet Research 14(1) (2012), doi:10.2196/jmir.1924
11. Wang, Y., Lin, J., Annavaram, M., Jacobson, Q.A., Hong, J., Krishnamachari, B., Sadeh, N.: A Framework of Energy Efficient Mobile Sensing for Automatic User State Recognition. In: Proceeding of MobiSys (2009)
12. Hong, Y., Kim, I., Ahn, S.C., Kim, H.: Mobile health monitoring system based on activity recognition using accelerometer. In: Simulation Modeling Practice and Theory, pp. 446–455 (2010)
13. Zappi, P., Stiefmeier, T., Farella, E., Roggen, D., Benini, L., Troster, G.: Activity recognition from on-body sensors by classifier fusion: sensor scalability and robustness. In: Proceeding of 3rd International Conference on Intelligent Sensors, Sensor Networks and Information, pp. 281–286 (2007)
14. Aziz, O., Atallah, B.L., ElHelw, M., Wang, L., Yang, G.Z., Darzi, A.: A Pervasive Body Sensor Network for Measuring Postoperative Recovery at Home. Surgical Innovation 14(2), 83–90 (2007)
15. Ermes, M., Parkka, J., Mantyjarvi, J., Korhonen, I.: The ingestible telemetric body core temperature sensor in Controlled and Uncontrolled Conditions. IEEE Transactions on Information Technology in Biomedicine 12(1), 20–26 (2008)
16. Kohonen, T.: The Self-Organizing Map. Proceedings of The IEEE 78(9), 1464–1480 (1990)
17. Chi, Z., Wu, J., Yan, H.: Handwritten numeral recognition using self-organizing maps and fuzzy rules. Pattern Recognition 28(1), 59–66 (1995)
18. Kitakyushu: SOM of SOMs. Neural Networks 22(4), 463–478 (2009)
19. Hu, W., Xie, D., Tan, T., Maybank, S.: Learning Activity Patterns Using Fuzzy Self-Organizing Neural Network. IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics 34(3), 1618–1626 (2004)
20. Pakkanen, J., Iivarinen, J., Oja, E.: The Evolving Tree – Analysis and Applications. IEEE Transactions on Neural Networks 17(3) (2006)
21. Vesanto, J., Alhoniemi, E.: Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks 11(3), 586–600 (2000)
22. Brugger, D., Bogdan, M., Rosenstiel, W.: Automatic Cluster Detection in Kohonen's SOM. IEEE Transactions on Neural Networks 19(3), 442–459 (2008)
23. Lau, K.W., Yin, H., Hubbard, S.: Kernel Self-Organsing Maps for Classification. Neurocomputing 69, 2033–2040 (2006)
24. Suganthan, P.N.: Hierarchical Overlapped SOM's for Pattern Classification. IEEE Transactions on Neural Networks 10(1), 193–196 (1999)
25. Li, Z., Eastman, J.R.: The Nature and Classification of Unlabelled Neurons in the Use of Kohonen's Self-Organizing Map for Supervised Classification. Transactions in GIS 10(4), 599–613 (2006)

26. Vesanto, J.: SOM-based Data Visualization Methods. Intelligent Data Analysis 3(2), 111–126 (1999)
27. Côme, E., Cottrell, M., Verleysen, M., Lacaille, J.: Aircraft Engine Health Monitoring Using Self-Organizing Maps. In: Perner, P. (ed.) ICDM 2010. LNCS (LNAI), vol. 6171, pp. 405–417. Springer, Heidelberg (2010)
28. Sirola, M., Lampi, G., Parviainen, J.: SOM Based Decision Support in Failure Management. In: IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, pp. 468–473 (2005)
29. Krause, A., Smailagic, A., Siewiorek, D.P.: Context-Aware Mobile Comupting: Learning Context-Dependent Personal Preferences from a Wearable Sensor Array. IEEE Transactions on Mobile Computing 5(2), 113–128 (2006)
30. Suzuki, S., Mitsukura, Y., Igarashi, H., Kobayashi, H., Harashima, F.: Activity recognition for children using self-organizing map. In: 2012 IEEE RO-MAN, pp. 653–658 (2012)
31. Hattori, Y., Kyushu, K., Inoue, S., Hirakawa, G.: Visualization for Activity Information Sharing System Using Self-Organizing Map. In: Proceeding of International Conference in Broadband and Wireless Computing, Communication and Applications (BWCCA), pp. 537–542 (2011)
32. Kurdthongmee, W.: A Self Organizing Map Based Motion Classifier with an Extension to Fall Detection Problem and Its Implementation on a Smartphone. Applications of Self-Organizing Maps (2012)
33. Seiffert, U.: Growing multi-dimensional self-organizing maps for motion detection. Self-Organizing Neural Networks (2002)
34. Ghasemzadeh, H., Barnes, J., Guenterberg, E., Jafari, R.: A Phonological Expression for Physical Movement Monitoring in Body Sensor Networks. In: Proceeding of 5th IEEE International Conference on Mobile Ad Hoc and Sensor Systems, pp. 58–68 (2008)
35. Bao, L., Intille, S.S.: Activity Recognition from User-Annotated Acceleration Data. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 1–17. Springer, Heidelberg (2004), doi:10.1007/978-3-540-24646-6_1

Range-Free Mobile Node Localization Using Static Anchor

Kaushik Mondal and Partha Sarathi Mandal

Department of Mathematics
Indian Institute of Technology, Guwahati, India
{mondal.k,psm}@iitg.ernet.in

Abstract. In this paper we have proposed a deterministic, range-free, distributed localization algorithm for mobile sensor nodes with static anchors. Mobile node calculates its approximate line of movement and corresponding position based on received beacons from two different anchors. The positional error can be further reduced by updating the approximate line of movement on receiving beacons from more anchors. We also have incorporated irregular radio propagation in our model. We have compared performance of our algorithm with existing localization algorithms. Simulation results show 80% improvement in performance of our proposed algorithm over the existing algorithms in terms of positional accuracy.

Keywords: Mobile sensor localization, Range-free, Beacon point, Line of movement.

1 Introduction

Mobile wireless sensor networks (MWSNs) is a recent development of wireless sensor networks (WSNs). There are lots of applications of MWSNs in service industry, house keeping, wildlife tracking, pollution monitoring, photon detection, shooter detection [8] *etc.* To run these applications successfully we need to look on the issues like coverage, localization, connectivity, energy consumption. The localization of mobile sensor nodes is most important for the above mentioned applications but localization of mobile sensor nodes with less error is a challenging problem.

GPS is commonly used technique for localizing mobile and static wireless devices. A few GPS enabled sensor nodes are usually deployed to localize mobile sensor nodes with consideration of cost and energy optimization in MWSN. Usually GPS enabled sensor nodes are termed as anchors, who know their own positions. Ideally communication area of any anchor is a circular disk in two dimension with the range of the anchor as its radius. However in practice, the radio propagation is usually not homogenous in all directions because of the presence of multi-path fading and different path losses depending on the direction of propagation, which is termed as radio propagation irregularity [19]. Due to irregular radio propagation, signal does not reach up to the boundary of the

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 269–284, 2013.

circular disk at every direction. According to the Fig. 4, if radius of the larger circle is the communication range, due to irregular radio propagation, sometimes signal reaches only up to the boundary of the smaller circle in some direction. Localization accuracy is affected due to irregular behavior of radio signal.

In MWSN, mobile sensor nodes are more powerful in terms of energy because they need to localize themselves frequently than static sensor nodes where localizing a static sensor node once is sufficient. Now a days, the developed mobile sensor nodes who can control their movement (for example, mobile actuated sensor [5,7]), are motivating researchers to find accurate localization methods. Generally there are three phases in a localization method, (1) coordination, (2) measurement or data gathering and (3) computation. Generally in MWSN, mobile nodes are used to record time stamp of events like receiving beacon in the coordination phase. This technique is used in many localization schemes [11]. Measurement phase is different for range-based and range-free algorithms. Range-based algorithms [12,14] depends on distant and angle measurements and generally produce better results than range-free. In this phase nodes gather information like hop count [15] in range-free algorithms. In the computation phase, approximate positions of the nodes are determined using the gathered information. Dead reckoning' [10,22] is a technique used in this phase for mobile node localization. In this technique the node calculates current position using previous position, moving speed and time difference between current time and the time when the position of the node was last updated. Mobile node localization methods can be centralized [11,12], as well as distributed. Since mobility requires rapid and continuous localization, distributed algorithms are more effective than the centralized algorithms.

1.1 Our Contribution

Using our algorithm, a mobile node can localize itself within a known error bound when it passes through communication circles of two different anchors. As it passes through more communication circles, positioning error can be further reduced. It is also possible to localize nodes within a predetermined error bound by fixing appropriate beacon distance. In our simulation study we have computed localization error considering irregular radio propagation of the anchors. Simulation results shows around 80% improvement of the positioning error over the existing algorithm [23]. After computing position of mobile node using proposed algorithm, 30% and 60% further reduction in error has been shown in simulation when node passes through three and five more communication circles respectively.

2 Related Works

Algorithms for localizing static sensor nodes can be applied for localizing mobile nodes, but computation cost is more since repeated run of the algorithm is needed as the mobile nodes change their position frequently. In the paper [18], authors Tilak et al. experimented how frequently localization algorithms for static nodes need to be executed to localize mobile nodes with an acceptable accuracy and energy consumption. Navstar global positioning system [9] is the mostly used technique for localizing mobile nodes. Kostas et al. proposed a range-based algorithm [3] for navigation of mobile robots. Datta et al. proposed an algorithm in [6] which can be used for both static and mobile sensor networks, where sensor nodes constructs polygon of presence and shrinks it using received information, while mobile nodes dilate it before sending to its neighbors. Ganggang et al. proposed an range based mobile node localization algorithm in [21], assuming that the mobile nodes are not always moving in the network. An algorithm proposed in [20], where mobile nodes predict their positions using recorded beacon information by guessing a mobility pattern under a statistical model. Hu et al. adapted monte carlo localization (MCL) in [10] to localize mobile nodes. Number of anchor nodes should be high to achieve good localization accuracy in this algorithm. Shafagh et al. proposed iCCA-MAP algorithm and compared it with MCL in [1]. Later Baggio et al. proposed monte carlo localization boxed (MCB) in [2], which improved on MCL by introducing anchor box to reduce the scope of presence of the node. Zhang et al. proposed weighted monte carlo localization (WMCL) in [23], which improved over MCB by reducing the size of anchor box (called bounding box here) constructed in MCB. This work is based on sequential monte carlo method. A range-free algorithm for mobile node localization is proposed in [16] depending on beacon point selection. The idea of modifying beacon point positions is good but in practice marking beacon points suffers from mobility of both nodes and anchors. There are also some range-free algorithms for localizing static sensor nodes using mobile anchors. Chia-Ho-Ou proposed a localization scheme in [4] using mobile anchors with directional antenna Ssu et al. proposed a range-free localization algorithm in [17], where mobile anchor nodes were used to find position of static sensor nodes. Later Lee et al. in [13] made improvement over that work by introducing some geometric constraints on the same model. Both these works are for static sensor nodes where anchors are mobile. Our proposed range-free algorithm for mobile sensor networks is motivated by the strategy of beacon point selection used in [13,17].

3 Basic Idea

Static anchors with equal communication range are deployed sparsely in a two dimensional plane. We can identify an anchor by its location. The anchors periodically broadcast beacons with their locations. The time interval between two consecutive broadcasts of beacons is t, which is fixed and same for all anchors. The mobile sensor nodes move according to the requirement of underlying application. During localization phase a node move with a uniform velocity until localization. When a mobile node receives first beacon from an anchor, it recognizes that it is in the communication range of that anchor. The position of receiving first beacon is an approximate end point of the chord of the communication circle of that anchor, along which it is moving. Similarly, the last beacon received from that anchor denotes the approximate position of the other end point of that chord. How to identify first and last beacons corresponding to an anchor is explained below. Beacons received at such points are marked and called as *beacon points*. Distance traveled by a mobile node between receiving any two consecutive beacons from an anchor is termed as *beacon distance* and which is equal to vt $(= u, say)$, where v is uniform velocity of the node in that localization phase. Mobile nodes know the transmission range of the anchors. So, when a mobile node moves into the transmission area of an anchor and receives a beacon, it knows the equation of the circle in which it is moving. We call the line along which a node is moving within a circle as its *actual line of movement*. The mobile nodes are equipped with digital compass for knowing the direction of movements and a timer to record the time stamps of the received beacons.

3.1 Beacon List

Each mobile node maintains a *beacon list* with two columns. Each entry of the list contains time stamp (*time_stamp*) of the received beacon and corresponding anchor id (*anchor_id*) according to its own clock. When a mobile node receives a beacon from an anchor, it records $< time\_stamp, anchor\_id >$ in the beacon list. At the beginning, list is empty. Let a node receives a beacon from an anchor with $anchor\_id = i$. It records $< time\_stamp, i >$ in the first column and marks it as a beacon point. After receiving each beacon from an anchor, node waits for a time greater than t for the next beacon from that anchor. If another beacon is received from the same anchor by that time then the node replaces the last beacon with latest one provided the last beacon is not marked as beacon point. If last beacon is marked as beacon point then records the latest beacon. In between whenever any beacon comes from a different anchor with $anchor\_id = j$ (say, $j \neq i$), it records $< time\_stamp, j >$ in the second column and marks it as a beacon point. If both the columns are non-empty and a beacon is received from an anchor with $anchor\_id = s$ $(s \neq i, j)$, then that beacon is ignored. Following the procedure discussed above, node records at most two beacon points for each anchor. There is possibility of receiving only one beacon from an anchor, in that case only one beacon point appears in a column of the list. So, there can be maximum of four entries in the list. The node deletes used marked beacon points and makes the list empty, after localizing itself as explained in section 4.2.

3.2 Finding Line of Movement

We are explaining here how a mobile node can identify its actual line of movement based on beacon points corresponding to two anchors. A mobile node calculates the time difference between receiving two beacon points from an anchor, to calculate the approximate chord length $2l$. Using the digital compass the node also knows the gradient of the line along which it moves. Now since there are only two chords of fixed length and fixed gradient in a circle, we can say the node is moving along any one of those two lines. For simplicity, let the node

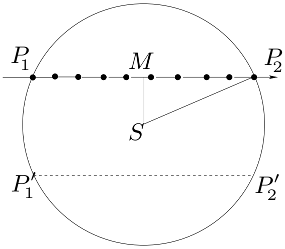


Fig. 1. Showing two possible lines of movement including the actual one (P_1P_2)

moves along a line parallel to x-axis Let one anchor be placed at $S = (0, 0)$ and P_1P_2 is the actual line of movement of the mobile node, where P_1, P_2 are the beacon points as shown in Fig. 1. Here the equation of the communication circle is $x^2 + y^2 = r^2$ and $SM = \sqrt{r^2 - l^2}$, where $P_1P_2 = 2l$, Hence, the possible line of movement of the mobile node is either P_1P_2 or $P_1'P_2'$, whose equations are $y = \pm\sqrt{r^2 - l^2}$. More information is needed to identify the actual line. For which the node continues its movement along the same line until it crosses the communication circle of another anchor. Let $S' = (a, b)$, $b \neq 0$ be a different anchor as shown in the Fig. 2. Let P_3, P_4 be the beacon points corresponding to S' and $S'M' = \sqrt{r^2 - l'^2}$, where $P_3P_4 = 2l'$. Similarly, the possible line of movement of the mobile node is either P_3P_4 or $P_3'P_4'$, whose equations are $y = b \pm \sqrt{r^2 - l'^2}$ Now, among these four lines, equations of two lines should be same with the actual line of movement, since there is exactly one line along which the node is moving.

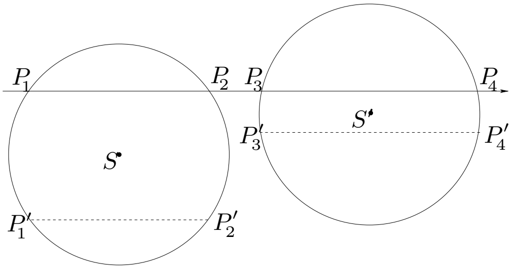


Fig. 2. Two communication circles help to find the correct line of movement

According to the Fig. 2, the equation of the actual line is $y = \sqrt{r^2 - l^2}$ *i.e.*, $y = b + \sqrt{r^2 - l'^2}$. Using the selected equation of the line of movement the node can calculate the coordinates of the points P_1, P_2, P_3, P_4. Current position of the mobile node can be calculated based on any P_i for $i = 1$ to 4, and elapsed time from timer. To localize itself, a node has to pass through communication circles of two different anchors where the line joining those two anchors should not be parallel with the line of the node, *i.e.*, $SS' \nparallel P_1P_4$.

3.3 Error Analysis

The above technique is not applicable for finding line of movement unless beacon points are located on the perimeter of the communication circles of the respective anchors. In practice beacon points, C, C' (ref. Fig. 3) may lie inside the communication circles due to periodic broadcast of anchors. In that case mobile node finds *approximate line of movement.* According to the Fig. 3, possible approximate lines of movement are $P_1'P_2'$ and $P_1''P_2''$. How they become the possible approximate lines of movement is discussed below. C, C' are two beacon points inside the communication circle of the anchor S. Mobile node can measure the length CC' from its velocity and time stamps of the beacon points. The node misinterprets the length CC' as the length of the chords P_1P_2 along which it is moving. So, the actual line of movement P_1P_2 is projected to the approximate line of movement $P_1'P_2'$ such that $CC' = P_1'P_2'$. Due to the symmetric nature of circle we will get another possible line of movement $P_1''P_2''$. It is possible to discard $P_1''P_2''$ as a line of movement using one more communication circle of another anchor as discussed later. The error in calculating line of movement is MM' as shown in Fig. 3. Now natural questing is coming when maximum possible error occurs. Following lemmas answer this query.

Lemma 1. *The positioning error is equal to the perpendicular distance between the actual line of movement and the approximate line of movement.*

Proof. According to the Fig. 3 the actual line of movement is P_1P_2 and the selected approximate line of movement is $P_1'P_2'$. Let T be the time stamp of receiving beacon at C'. With the knowledge of $P_1'P_2'$ the mobile node calculates position of the beacon point as P_2' instead of C'. Node also calculates the distance traveled by itself after receiving that beacon point using the current time from its timer, which is $C'Q$. Since $C'Q$ is equal to $P_2'Q'$, the error in positioning is equal to QQ', which is equal to the perpendicular distance between the actual line of movement and the approximate line of movement. □

Lemma 2. *The value of the possible error is maximum when mobile node passes along the diameter of the communication circle of any anchor.*

Proof. According to the Lemma 1, positioning error is equal to QQ'. $QQ' = MM'$ as shown in the Fig. 3. Hence, maximum possible value of MM' for all possible chords is the maximum possible error in localization.

As the values of P_1C and P_2C' increases, MM' also increases. Without considering the irregularity in signal propagation $[0, u)$ is the range of possible values of P_1C and $C'P_2$. When $P_1C = u = C'P_2$, then the length of the actual chord along which the node is moving becomes $2l+2u$, where $2l$ is the calculated chord length. So, length of $MM' = Err(2l) = \sqrt{r^2 - l^2} - \sqrt{r^2 - (l+u)^2}$. For any fixed u, it gives maximum value when $u + l = r$. That is, if the node moves along the diameter and calculates the chord length as $2(r-u)$ instead of $2r$, then the error becomes maximum. □

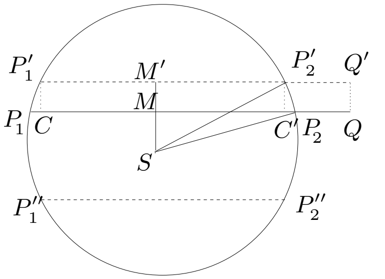


Fig. 3. Showing maximum error MM' for the line of movement P_1P_2

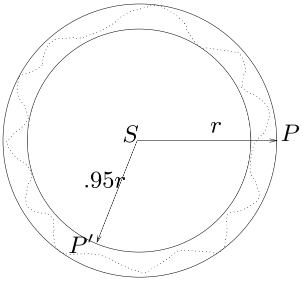


Fig. 4. SP and SP' are radii of the maximum and minimum possible communication range of S due to effect of irregular radio propagation where $SP - SP' = .05r$

We denote ϵ as the maximum possible error and $\epsilon = \sqrt{r^2 - (r-u)^2}$. From this relation, fixing a value for beacon distance u is possible such that the maximum possible error ϵ stays within any predetermined value. For irregular radio propagation, the range of possible values of P_1C and $C'P_2$ in Fig. 3 belongs to $[0, u + 0.01kr)$, where k is the percentage of maximum possible reduction of the communication range due to radio propagation irregularity as shown in the Fig. 4. Expression of the maximum possible error becomes $\epsilon = \sqrt{r^2 - (r - u - .01kr)^2}$. So, we can keep ϵ under control by choosing appropriate values of u and r. For our simulation study we have consider $k = 5$ in section 6. However, if it is not known, we can estimate it by sending radio signals and measuring the maximum and minimum communication ranges of environments.

An interesting observation is that, if there is any error while selecting the approximate line of movement, the erroneous line is always far from the anchor than the actual line of movement. It happens because the length of the chord always decreases in case of wrong chord length estimation. This fact helps us while choosing the approximate line of movement among four equations corresponding to two anchor's communication circle as explained below. As discussed in section 3.2, a node chooses the actual line of movement among four lines calculated by passing through two communication circle of different anchors in ideal situation, when beacon points are on the perimeter of the circles. But the following result ensures that the node chooses the approximate line of movement with error bound ϵ in practical situation, when beacon points are inside the circles.

Theorem 1. *Among four equations of possible lines of movements, a mobile node finds the pair of line such that perpendicular distance between them is the least. If the node chooses any one line from that pair then error will always be at most ϵ, where perpendicular distance between the lines passing through the anchors with same gradient with the actual line of movement is at least ϵ.*

Proof. For simplicity, we prove the theorem assuming that the mobile node is moving along a line parallel to x-axis. There are two possible cases:

Case 1: If actual line of movement passes through same side of two anchors. According to Fig. 5 the actual line of movement is P_1P_4, passes through upper side of the anchors S and S'. Let coordinates of S and S' be (a, b) and (c, d) respectively, where $(d - b) \geq \epsilon$. Let equation of the actual line of movement be

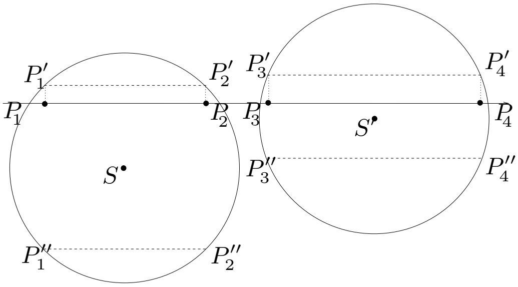


Fig. 5. Anchors are at same side of the line of movement

$$y = dk \text{ where } k \geq 1 \tag{1}$$

Let the chord length corresponding to the communication circles of the anchors S and S' be $2l$ and $2l'$ respectively. According to Lemma 2, node calculates equation of the line $P'_1P'_2$ and $P''_1P''_2$ as,

$$y = dk + \epsilon/n_1 \text{ where } n_1 \geq 1 \text{ and} \tag{2}$$

$$y = 2b - dk - \epsilon/n_1 \text{ where } n_1 \geq 1 \tag{3}$$

respectively. Similarly, node calculates equation of the line $P'_3P'_4$ and $P''_3P''_4$ as

$$y = dk + \epsilon/n_2 \text{ where } n_2 \geq 1 \text{ and} \tag{4}$$

$$y = 2d - dk - \epsilon/n_2 \text{ where } n_2 \geq 1 \tag{5}$$

respectively. The distance between the lines (Eqn. 2 and Eqn. 4) is equal to $|\epsilon/n_2 - \epsilon/n_1| \leq \epsilon$. We denote distance between lines (Eqn. i) and (Eqn. j) by $D(i, j)$ hereafter. So, $D(2, 4) \leq \epsilon$. Let it be the smallest distance among those six possible distances $D(i, j)$. If we choose Eqn. 2 or Eqn. 4, we are choosing lines within ϵ error, since $D(1, 2) = \epsilon/n_1 \leq \epsilon$ as well as $D(1, 4) = \epsilon/n_2 \leq \epsilon$. Actually we are finding distances between correct line (Eqn. 1) and the line we are choosing according to our technique and checking it is less or equal to ϵ or not. Eqn. 2 and Eqn. 4 make sure that there is at least one pair of lines whose perpendicular distance is less than or equal to ϵ.

Let $D(2, 3)$ be the least. We have to check $D(1, 3)$ only. $D(1, 3) = 2dk - 2b + \epsilon/n_1 \leq 2dk - 2b + 2\epsilon/n_1 = D(2, 3) < D(2, 4) = |\epsilon/n_2 - \epsilon/n_1| \leq \max(\epsilon/n_1, \epsilon/n_2)$.

Let $D(2,5)$ be the least. We have to check $D(1,5)$ only. $D(1,5) = 2dk - 2d + \epsilon/n_2 \leq (2dk - 2d + \epsilon/n_1 + \epsilon/n_2) = D(2,5) < D(2,4) = |\epsilon/n_2 - \epsilon/n_1| \leq \max(\epsilon/n_1, \epsilon/n_2)$.

$D(3,4) = 2dk - 2b + \epsilon/n_1 + \epsilon/n_2 > 2dk - 2d + \epsilon/n_1 + \epsilon/n_2 = D(2,5)$, since $d > b$. So, it cannot be the least.

$D(3,5) = 2d - 2b + \epsilon/n_1 - \epsilon/n_2 \geq \epsilon$, since $d \geq (b + \epsilon)$. So it cannot be the least.

Let $D(4,5)$ be least. We have to check $D(1,5)$ only. $D(1,5) = 2dk - 2d + \epsilon/n_2 \leq 2dk - 2d + 2\epsilon/n_2 = D(4,5) \leq |\epsilon/n_2 - \epsilon/n_1| \leq \max(\epsilon/n_1, \epsilon/n_2)$.

Hence we show that if $D(i,j)$ is the least and we choose any one of the $i-$th and $j-$th equation, then we are choosing a line of movement which is within $\max(\epsilon/n_1, \epsilon/n_2)$ error with the actual line of movement.

Case 2: If actual line of movement passes through different sides of two anchors.

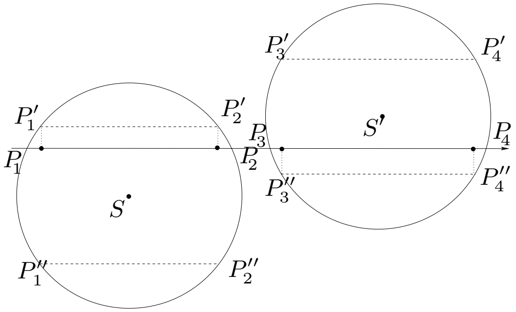


Fig. 6. Anchors are at different sides of the line of movement

According to Fig. 6, the actual line of movement is P_1P_4 passes through the different sides of S and S'. Let the equation of the actual line of movement be

$$y = d/k \text{ where } k \geq 1 \tag{6}$$

Let the chord length corresponding to the communication circles of the anchors S and S' be $2l$ and $2l'$ respectively. According to Lemma 2, node calculates equation of the line $P'_1P'_2$ and $P''_1P''_2$ as,

$$y = d/k + \epsilon/n_1 \text{ where } n_1 \geq 1 \text{ and} \tag{7}$$

$$y = 2b - d/k - \epsilon/n_1 \text{ where } n_1 \geq 1 \tag{8}$$

respectively. Similarly, node calculates equation of the line $P'_3P'_4$ and $P''_3P''_4$ as,

$$y = d/k - \epsilon/n_2 \text{ where } n_2 \geq 1 \text{ and} \tag{9}$$

$$y = 2d - d/k + \epsilon/n_2 \text{ where } n_2 \geq 1 \tag{10}$$

respectively. Now perpendicular distance between Eqn. 7 and Eqn. 9 is equal to $|\epsilon/n_2 + \epsilon/n_1| \leq 2\epsilon$. So, $D(7,9) = \epsilon/n_1 + \epsilon/n_2 \leq 2\epsilon$. Let it be the smallest distance among those six possible distances. If we choose Eqn. 7 or Eqn. 9, we are choosing lines within ϵ error, since $D(6,7) \leq \epsilon$ as well as $D(6,9) \leq \epsilon$. Eqn. 7 and Eqn. 9 make sure that there is at least one pair of lines whose perpendicular distance is less than or equal to 2ϵ.

Let $D(7,8)$ be least. We have to check $D(6,8)$ only. $D(6,8) = 2d/k - 2b + \epsilon/n_1 \leq 2d/k - 2b + 2\epsilon/n_1 = D(7,8) < \epsilon/n_1 + \epsilon/n_2 = D(7,9)$, implies, $D(6,8) = 2d/k - 2b + \epsilon/n_1 \leq \epsilon/n_2$.

Let $D(7,10)$ be least. We have to check $D(7,10)$ only. $D(7,10) = 2d/k - 2d - \epsilon/n_2 \leq 2d/k - 2d + \epsilon/n_1 - \epsilon/n_2 = D(7,10) < \epsilon/n_1 + \epsilon/n_2 = D(7,9)$, implies, $D(6,10) = 2d/k - 2d - \epsilon/n_2 \leq \epsilon/n_2$.

$D(8,9) = 2d/k - 2b + \epsilon/n_1 + \epsilon/n_2 > 2d/k - 2d + \epsilon/n_1 + \epsilon/n_2 = D(7,10)$, since $d > b$. So, it cannot be the least.

$D(8,10) = 2d - 2b + \epsilon/n_1 - \epsilon/n_2 \geq 2\epsilon$, since $d \geq (b + \epsilon)$. So it cannot be the least.

Let $D(9,10)$ be least. We have to check $D(7,10)$ only. $D(7,10) = 2d/k - 2d + \epsilon/n_2 \leq 2d/k - 2d + 2\epsilon/n_2 = D(9,10) < D(7,9) = \epsilon/n_1 + \epsilon/n_2$, implies, $D(6,10) = 2d/k - 2d + \epsilon/n_2 \leq \epsilon/n_1$.

Hence in this case also, if $D(i,j)$ is the least and we choose any one of the $i-$th or $j-$th equation, then we are choosing a line of movement which is within $\max(\epsilon/n_1, \epsilon/n_2)$ error with the actual line of movement. □

Since the calculated chord length are $2l$, $2l'$ and the error function depends on the chord length, so $\max(\epsilon/n_1, \epsilon/n_2)$ is equal to $\max(Err(2l), Err(2l'))$ in the worst case. Since we saw earlier that the possible error ϵ is achieved when the chord length is the maximum, so error decreases as the perpendicular distance of the chord from the anchor increases. That is, the value of maximum possible error is less when the node is far from the anchor.

4 Proposed Localization Algorithm

4.1 System Model

Static anchors are located sparsely over a mobile sensor network on a two dimensional plane. The anchors are GPS enabled. They know their own positions and are uniquely identified by their positions. The anchors broadcast beacons with their positions periodically. Mobile sensor nodes are moving with a uniform velocity during localization and receive beacons when they are in transmission ranges of any anchor. The transmission range of the anchors are known to the mobile nodes. In case of irregular radio propagation maximum and minimum possible transmission ranges are also known to the mobile nodes. All mobile nodes are attached with a timer and a digital compass so that they can calculate traveled distance and direction of movement. A mobile node does not change direction during localization.

4.2 Position Calculation

When a mobile node finds all marked beacon points in its beacon list corresponding to two anchors, it calculates the time difference in the time stamps of two beacon points corresponding to an anchor for calculating approximate chord length of the communication circle of the same anchor. Calculate two equations of possible lines according to the chord length. Similarly, it calculates two equations of possible lines for the second anchor. If there is only one beacon point corresponding to an anchor then the possible lines are tangents since the calculated chord length is zero. According to the theorem 1, find the pair of lines with minimum distance between them. Our technique selects the line by the following rule. Take the communication circle whose anchor has lesser y-coordinate and find the chord lengths using those two equations. Select that line which has larger chord length on the communication circle of the anchor, as the approximate line of movement. Find the time stamp of the last beacon received by the node from any one of those anchors. Find the coordinate of the intersection point of the communication circle of the same anchor and the selected equation of line. Since it is moving along the same line until localization, using the current time in its clock find its current position. The node removes used beacon point entries from the beacon list. Once a node localizes itself, if it keep tracks of its direction changes and corresponding time using the directional antenna, then it can calculate its position at any time instance.

4.3 The Algorithm

A mobile node calculates its position using the following MobileNodeLocalization algorithm. Appropriate expression of ϵ should be chosen depending on the presence of radio propagation irregularity or not as described in section 3.3.

Algorithm 1: MobileNodeLocalization

```
Mobile node begins localization process by moving with a uniform
velocity until localization.
Maintains beacon list as described in section 3.1.
if there are four marked beacon points in beacon list such that the
distance between the lines, which are passing through the anchors
parallel to the direction of the node's movement is at least $\epsilon$ then
    the node calculates the current position according to section 4.2.
end if
```

5 Error Minimization

Sensor node localizes itself using the above algorithm within error bound ϵ. There is a possibility of reducing error by updating the line of movement whenever it passes through communication circle of some anchor as explained below. Suppose a node computed its approximate line of movement $y = mx + c$ (say), by the above algorithm. Currently the node is passing through the communication circle

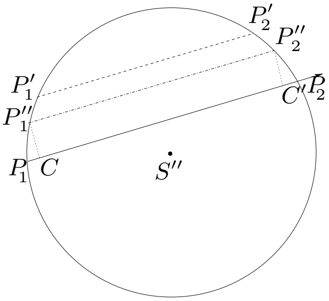


Fig. 7. Node minimizes error by updating line of movement

of an anchor S'', where the actual line of movement is P_1P_2 and the approximate line of movement is $P_1'P_2'$ as shown in the Fig. 7.

Error can be reduced by finding a suitable approximate line of movement between $P_1'P_2'$ and P_1P_2. To do this, the node approaches as follows: The node calculates length l (say) of $P_1'P_2'$ from the equations of the circle and the line $y = mx + c$. The node also calculates length L (say) of the chord along which it is moving using two beacon points C, C' corresponding to the anchor S''. If it finds $L > l$, error reduces by updating the approximate line of movement $P_1'P_2'$ to $y = mx + c'$, which is equation of the chord of length L, where $c' = \left[ma - b \pm \sqrt{(}(4r^2 - l^2)(m^2 + 1))/2\right]$ and (a, b) is the position of S''. Among these two lines $y = mx + c'$ (for different values of c'), which is closer to $P_1'P_2'$ is chosen as the new approximate line of movement. According to the Fig. 7, $P_1''P_2''$ is the updated line. If the node finds $L \leq l$ no improvement in error is possible, hence it does not change the approximate line of movement.

6 Simulation Results

We have used MATLAB platform to study the performances of our proposed scheme. We have done it under different communication ranges and beacon distances. We have generated random straight lines as the equation of actual lines of movement corresponding to two communication circles of distinct anchors, such that the distance between the lines passing through those two anchors with same direction as of the node is at least ϵ. To be more practical, we have used irregular radio model for simulation. Table 1 shows the average error without considering irregular radio model with respect to different communication ranges as well as beacon distances. The average error is very much less than the maximum possible error under irregular radio model, which makes the algorithm very effective. It also shows that the average error under irregular radio model increases compared to the average error without radio irregularity. Fig. 8 supports the table. In the graph, communication range is along x-axis and error is along y-axis. Here beacon distance is equal to $r/10$ and $k = 5$ due to radio propagation irregularity as discussed in section 3.3. For each beacon distance (BD), there are

Table 1. Showing values (in meter) of average error (AE), average error under radio irregularity (AERI) and maximum possible error under radio irregularity (MERI) for different communication range (CR) and beacon distance (BD)

BD→	**1**			**3**			**5**		
CR↓	AE	AERI	MERI	AE	AERI	MERI	AE	AERI	MERI
10	0.82	1.18	5.26	1.84	1.91	7.59	2.47	2.62	8.93
20	1.03	1.77	8.71	2.26	2.83	12.00	3.24	3.86	14.28
30	1.17	2.27	11.98	2.57	3.50	15.80	3.68	4.35	18.64
40	1.19	3.01	15.19	3.07	4.10	19.36	4.01	5.18	22.60
50	1.27	3.29	18.37	2.94	4.65	22.79	4.54	5.67	26.33

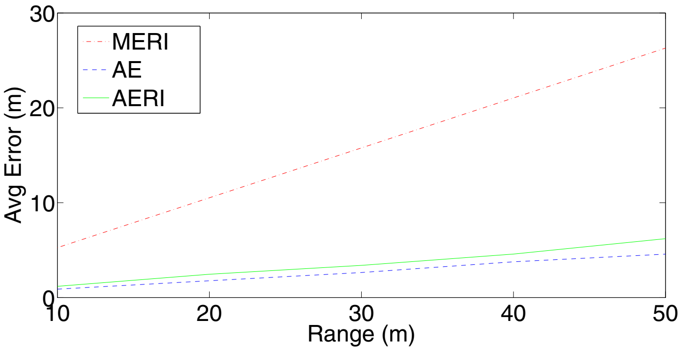


Fig. 8. Error increases under irregular radio model but much lesser than maximum possible error

three columns in the Table 1 showing the values of average error (AE), average error under radio irregularity (AERI) and maximum possible error under radio irregularity (MERI). All data are in meter.

As we discussed in the section 5, that node can minimize the localization error as it passes through the communication circles of other anchors. Results shows that, node can minimize error with a good efficiency. Following Table 2 shows the updated average error as the mobile node passes through three more communication circles (AE3) and five more communication circles (AE5) after localization. For each beacon distance, AE3 and AE5 are shown in two columns.

In the Fig. 9 communication range is along x-axis and average error is along y-axis. Here beacon distance is equal to $r/10$ and $k = 5$ due to radio propagation irregularity as discussed in section 3.3. Fig. 9 shows significant improvement of the positioning error when node passes through three and five more communication circles respectively for different communication ranges. In Table 3, we have compared our result with some of the existing mobile node localization algorithms WMCL-A and WMCL-B [23], MCL [10], MCB [2]. We have taken the communication range as 20 meter, beacon distance 2 meter and radio irregularity as discussed before. Among those, WMCL-A and WMCL-B give best results.

Table 2. Showing values (in meter) of average error after passing through three more communication circles (AE3) and five more communication circles(AE5) under irregular radio range for different communication range (CR) and beacon distance (BD)

BD →	**1**		**2**		**3**		**4**		**5**	
CR↓	AE3	AE5	AE3	AE5	AE3	AE5	AE3	AE5	AE3	AE5
10	0.54	0.38	0.78	0.57	0.95	0.75	1.21	0.85	1.32	1.02
20	0.82	0.65	1.11	0.80	1.33	0.99	1.58	1.12	1.82	1.35
30	1.07	0.82	1.36	0.98	1.66	1.21	1.78	1.49	2.09	1.57
40	1.26	1.04	1.70	1.22	2.06	1.40	2.16	1.67	2.55	1.68
50	1.54	1.14	1.93	1.38	2.15	1.69	2.47	1.92	2.70	2.12

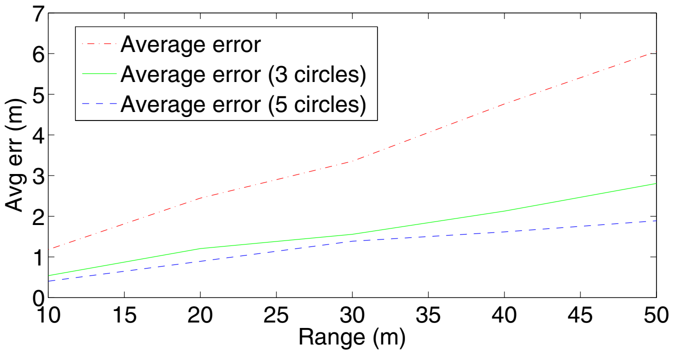


Fig. 9. Error decreases after passing through three and five communication circles

These two are sequential monte carlo analysis based algorithms. Simulation results shows around 80% improvement over WMCL-A of the positioning error as shown in Table 3. As the node moves along the network, after passing through three and five more communication circles, the error further reduces to $0.07r$ and $0.04r$ respectively from $0.1015r$. So error reduces around 30% and 60% after passing through three and five more communication circles respectively.

Table 3. Average error (r) of our proposed algorithm compared with some existing algorithms, where r is the communication range of anchor

Algorithm	Proposed algorithm	WMCL-A	WMCL-B	MCB	MCL
Error(r)	0.1015	0.5034	0.5683	0.6628	0.7669

7 Conclusion

In this paper we have proposed a deterministic range-free localization algorithm for mobile wireless sensor networks under irregular radio model. We have given bounds on errors in localization. Our algorithm is able to localize nodes within any predetermined error bound by fixing appropriate beacon distance. Simulation results show that in our approach, average error is very less than the maximum possible error and the algorithm outperforms over the existing approaches in terms of positional accuracy. Using our technique a mobile node can further reduces localization error whenever it passes through the communication circle of different anchors by repeated calculation of the approximate line of movement. Future work includes generalization of the localization algorithm under more practical scenario.

Acknowledgement. The first author is thankful to the Council of Scientific and Industrial Research (CSIR), Govt. of India, for financial support during this work.

References

1. Alikhani, S., Kunz, T., St-Hilaire, M.: ICCA-MAP versus MCL and dual MCL: Comparison of mobile node localization algorithms. In: Nikolaidis, I., Wu, K. (eds.) ADHOC-NOW 2010. LNCS, vol. 6288, pp. 163–176. Springer, Heidelberg (2010)
2. Baggio, A., Langendoen, K.G.: Monte-carlo localization for mobile wireless sensor networks. In: Cao, J., Stojmenovic, I., Jia, X., Das, S.K. (eds.) MSN 2006. LNCS, vol. 4325, pp. 317–328. Springer, Heidelberg (2006)
3. Bekris, K.E., Argyros, A.A., Kavraki, L.E.: Angle-based methods for mobile robot navigation: Reaching the entire plane. In: ICRA, pp. 2373–2378 (2004)
4. Ou, C.-H.: A localization scheme for wireless sensor networks using mobile anchors with directional antennas. IEEE Sensors Journal 11(7), 1607–1616 (2011)
5. Dantu, K., Rahimi, M.H., Shah, H., Babel, S., Dhariwal, A., Sukhatme, G.S.: Robomote: enabling mobility in sensor networks. In: IPSN, pp. 404–409 (2005)
6. Datta, S., Klinowski, C., Rudafshani, M., Khaleque, S.: Distributed localization in static and mobile sensor networks. In: WiMob, pp. 69–76 (2006)
7. Friedman, J., Lee, D., Tsigkogiannis, I., Wong, S., Chao, D., Levin, D., Kaiser, W., Srivastava, M.: RAGOBOT: A new platform for wireless mobile sensor networks. In: Prasanna, V.K., Iyengar, S.S., Spirakis, P.G., Welsh, M. (eds.) DCOSS 2005. LNCS, vol. 3560, p. 412. Springer, Heidelberg (2005)
8. Fuller, R., Koutsoukos, X.D. (eds.): MELT 2009. LNCS, vol. 5801. Springer, Heidelberg (2009)
9. Hofmann-Wellenhof, B., Lichtenegger, H., Collins, J.: Global Positioning System: Theory and Practice, 4th edn. Springer (1997)
10. Hu, L., Evans, D.: Localization for mobile sensor networks. In: MOBICOM, pp. 45–57 (2004)
11. Kusy, B., Lédeczi, Á., Koutsoukos, X.D.: Tracking mobile nodes using RF doppler shifts. In: SenSys, pp. 29–42 (2007)
12. Kusy, B., Sallai, J., Balogh, G., Lédeczi, Á., Protopopescu, V.A., Tolliver, J., DeNap, F., Parang, M.: Radio interferometric tracking of mobile wireless nodes. In: MobiSys, pp. 139–151 (2007)

13. Lee, S., Kim, E., Kim, C., Kim, K.: Localization with a mobile beacon based on geometric constraints in wireless sensor networks. IEEE Transactions on Wireless Communications 8(12), 5801–5805 (2009)
14. Niculescu, D., Badrinath, B.R.: Ad hoc positioning system (APS) using AOA. In: INFOCOM (2003)
15. Niculescu, D., Nath, B.: DV based positioning in ad hoc networks. Telecommunication Systems 22(1-4), 267–280 (2003)
16. Ou, C.-H.: Range-free node localization for mobile wireless sensor networks. In: ISWPC, pp. 535–539 (2008)
17. Ssu, K.-F., Ou, C.-H., Jiau, H.C.: Localization with mobile anchor points in wireless sensor networks. IEEE Trans. on Vehicular Technology 54(3), 1187–1197 (2005)
18. Tilak, S., Kolar, V., Abu-Ghazaleh, N.B., Kang, K.-D.: Dynamic localization protocols for mobile sensor networks. CoRR, cs.NI/0408042 (2004)
19. Velimirovic, A.S., Djordjevic, G.L., Velimirovic, M.M., Jovanovic, M.D.: Fuzzy ring-overlapping range-free (FRORF) localization method for wireless sensor networks. Computer Communications 35(13), 1590–1600 (2012)
20. Yi, J., Koo, J., Cha, H.: A localization technique for mobile sensor networks using archived anchor information. In: SECON, pp. 64–72 (2008)
21. Yu, G., Yu, F., Feng, L.: A localization algorithm using a mobile anchor node under wireless channel. In: ROBIO, pp. 1104–1108 (2007)
22. Zhang, P., Martonosi, M.: LOCALE: Collaborative localization estimation for sparse mobile sensor networks. In: IPSN, pp. 195–206 (2008)
23. Zhang, S., Cao, J., Chen, L., Chen, D.: Accurate and energy-efficient range-free localization for mobile sensor networks. IEEE Trans. Mob. Comput. 9(6), 897–910 (2010)

Neighbor Discovery Algorithm Based on the Regulation of Duty-Cycle in Mobile Sensor Network

Jinbao Li[1,2], Jian Yang[1,2], Yanqing Zhang[1,2],
Longjiang Guo[1,2], and Yingshu Li[3]

[1] School of Computer Science and Technology, Heilongjiang University
Harbin, Heilongjiang, China, 150080
[2] Key Laboratory of Database and Parallel Computing of Heilongjiang Province
Harbin, Heilongjiang, China, 150080
[3] Department of Computer Science, Georgia State University
Atlanta, Georgia, USA, 30303

Abstract. In this paper, aiming at the problem that the node discovers its neighbor nodes, when it is moving in mobile sensor network, we propose an algorithm for dynamical regulating the duty cycle of the node which needs to discover it neighbor node. This algorithm uses the boundary nodes in the communication range of the node to predict potential neighbor nodes, applies passion point process to predict the number of nodes in communication range of the node, regulates the duty-cycle of the node based on the model of balls and boxes, and then schedules the waking time of the node periodically by using the duty-cycle which is regulated to finish the detection. Finally, theoretical analysis and simulation experiments show that the proposed algorithm can discover more neighbor nodes in the short period of time with less energy.

Keywords: Mobile wireless sensor networks, sleep-wake scheduling, duty cycle, neighbor node.

1 Introduction

With the continuous development of electronic equipment, its sizes is getting smaller and smaller, and the electronic device arranged in a fixed location becomes the portable electronic device, which makes the static sensor networks also gradually become mobile sensor networks. In recent years, mobile sensor networks have become a research focus. Mobile sensor networks have a wide range of applications, such as military defense, animal detection, and social network, etc. In these applications, with the change of the mobile environment, the mobile sensor nodes will rediscover their neighbor nodes, and forward their effective data to these nodes. Because of sensor node powered by batteries, the node energy is limited, so the nodes cannot have a long time to discover neighbor in order to save energy to prolong the network lifetime.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 285–299, 2013.

In sensor networks, sleep scheduling mechanism is the energy-saving method which is most commonly used by the nodes. Each node applies the setting duty-cycle to work, discoveries neighbors, collects data in the wake-up time, and saves energy during sleep time. However, with the node moving, its neighbor nodes will increase or decrease and the node frequently requires to be woken up to discover neighbor nodes, which will consume energy. The traditional neighbor discovery algorithms apply a broadcast to complete neighbor discovery, and the nodes are woken up and detect all the neighbor nodes within a cycle time, which leads to great energy consumption. With the movement of the nodes, it constantly has new neighbor nodes entering their communication range, and when the nodes run faster, the nodes need to update the neighbor nodes and frequently monitor, therefore the energy of the nodes becomes a bottleneck. The nodes can also apply the time synchronization to complete neighbor node discovery, but the price of the time synchronization is larger.

In mobile sensor networks, due to the restrictions of movement speed and energy, it makes the neighbor discovery difficult to realize. The existing neighbor discovery algorithms are mostly passive discovery algorithms, which are prone to unstable discovery delay. However, among the applications, when the nodes move, they need to quickly discovery neighbor nodes in order to guarantee the reliability and real-time of the applications. The passive algorithms cannot meet the actual demand, therefore we need to design efficiently active neighbor discovery algorithm for mobile sensor networks. In order to make the nodes use less energy to discovery more neighbor nodes in the process of moving, we provide a Neighbor discovery algorithm Based on the Regulation of Duty-cycle in mobile sensor network (NBRD for short). The algorithm applies the neighbor nodes within the boundaries of the original communication range to predict the potential neighbor nodes and estimate the number of nodes in the new communication range. Then the algorithm takes advantage of balls and boxes model to dynamically adjust the duty cycle of the neighbor nodes, and modify the node wake-up time in wake cycle to discover neighbor nodes, which makes the node use less energy to quickly discover more neighbors.

The rest of this paper is organized as follows. Section 2 describes the related work. Section 3 surveys the assist mechanism of the boundary nodes. Section 4 introduces the F-node duty cycle adjustment mechanism, followed by its analysis in Section 5. Section 6 evaluates the NBRD's performance in simulation experiments. Section 7 concludes the paper.

2 Related Work

Currently, many researchers have studied on the neighbor discovery. For the static networks, aiming at the problem of neighbor discovery for wireless adhoc networks, J.McGlynn *et al.* provided the Birthday protocol[1]. This protocol applied the birthday paradox theory to calculate the slot allocation of wake, listening, sending, so that the communication links can be found with high probability. For mobile sensor networks, D.Culler *et al.* proposed Disco algorithm

[2] and Arvind *et al.* proposed a U-Connect algorithm [3]. These algorithms mainly utilized the idea of the Chinese remainder theorem to solve the problem of neighbor discovery. However, Disco and U-Connect algorithms belong to passive discovery algorithm by collision in the process of moving through the initially-set duty cycle, and they cannot guarantee that the nodes found more neighbors in a shorter time.

Tseng *et al.* proposed Quorum algorithm [4]. Quorum algorithm needs to ensure that all nodes have the same duty cycle and are the lack of flexibility. Aiming that Quorum algorithm used the same duty, the literature [5] provided a heterogeneous Quorum algorithm, and this algorithm allowed a node using a different duty cycle in order to save energy consumption. Because the neighbor discovery algorithms for Single-Radio Single-Channel networks do not apply to Single-Radio Multi-Channel networks, Neils *et al.* converted the general optimization problem into a linear programming problem, and proposed the solution of a linear programming, which mainly had collected information to shorten the neighbor discovery time and was a passive discovery algorithm [6]. The U-connect and Disco algorithms affect the average network performance, and deterministic algorithms [2] and probabilistic algorithms [1] will lead to a long discovery time. For these two types of problems, the literature [7] gave a Searchlight solution,which is a $\sqrt{2}$-approximate optimal algorithm. Aiming at confliction in the process of neighbor discovery, the literature [8] proposed ALOHA algorithm with conflict detection. This algorithm firstly transformed neighbor discovery into coupon collection problem, then calculated the time expectation to discover all nodes, according to the characteristics of coupons collection.

Aveek *et al.* proposed a WiFlock protocol for mobile sensor network, which combined neighbor discovery and group maintenance [10]. Although it had a better rate of discovery, the synchronization listening consumes more energy. Meanwhile, the protocol is committed to group maintenance, and network performance is susceptible to the change of the node density. In order to avoid confliction caused by the node density increase to affect the performance of the protocols, the literature [11] proposed a multichannel neighbor discovery protocol OPT and SWOPT. The protocol utilized linear programming to solve the node sleep scheduling and channel scheduling. However, the protocol is not suitable for the energy-constrained sensor network. Although the proposed protocol [12] also has a higher efficiency, similarly it is not applied to sensor networks with limited resources and energy.

The above researches about neighbor discovery mostly are passive discovery algorithms, and initially set the duty cycle of each node by some way. It will lead to the instability of discovery delay.The proposed dynamic adjustment mechanism actively regulates the duty cycle of the node which discovers neighbor nodes, according to the expected number of the undiscovered nodes in the new communication area, which avoids the instability of passive algorithm and is suitable for the sensor network.

3 Boundary Node Assist Mechanism

The mechanism makes full use of the boundary node to assist nodes to discover the potential neighbor nodes.

3.1 System Model

Definition 1: The node which needs to actively engage in the neighbor discovery is defined as F-node. Assumed that the distribution of nodes follows of Poisson distribution, and there is an F-node *Dis* that needs neighbor discovery. The Communication radius of all nodes is R.In the network initialization, the neighbor nodes in the communication radius of *Dis* are known. Let $Boundary\_Set$ denote the boundary node set of *Dis*, and $|Boundary\_Set|$ is the number of the boundary nodes.

After *Dis* moves a time period t, the angle between the original direction and the moving direction is θ, shown in Fig. 1. Let $top-\theta$, $top-2\theta$ denote the neighbor node sets. The nodes are in $top-\theta$ are $\omega_1 = |Boundary\_Set| * \theta/2\pi$ neighbor nodes closest to *Dis*, and the nodes are in $top-2\theta$ are $\omega_2 = |Boundary\_Set| * \theta/\pi$ neighbor nodes closest to *Dis*. The $top-\theta$ and $top-2\theta$ are used to monitor the boundary nodes.

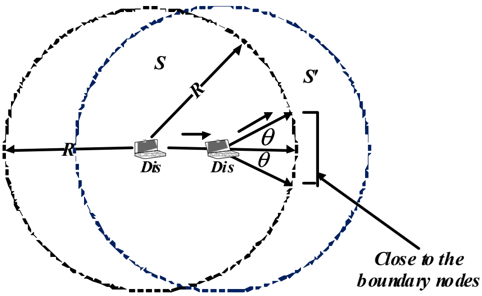


Fig. 1. Node movement diagram

In Fig. 1, the intersection region S of two communication circles is unchanged communication region, which contains the original neighbor nodes. The region S is a newly added communication region, containing newly added neighbors. Let *distD* denote the displacement produced by *Dis* moving.

$$dist_D = \sum_{i=1}^{T} V_{dif} * Time_i$$

Where V_{dif} is a variable of different speed, T is the time variable, $Time_i$ represents whether the speed used at the present time is V, and if not, $Time_i$ is 0, otherwise 1.

3.2 Boundary Nodes Assist Scheduling

Suppose the wake-up cycle of Dis is T, the assist scheduling scheme of its boundary nodes is to the set of $top - \theta$ as an example as follows.

1. Evenly divide a wake-up cycle time T into the nodes in $top - \theta$ and take T/ω_1 as the wake up time.
2. The wake-up time of the nodes in $top - \theta$ is W_1, W_2 ,W_3 ,..., W_{ω_1}, respectively, where $W_1 \cap W_2 \cap W_3 \cap \ldots \cap W_{\omega_1} = \Phi$, namely, the wake-up time of each node does not overlap.

When the boundary nodes in $top - \theta$ continuously are woken up in a wake-up cycle, it will find all of potential neighbors.

4 F-Node Duty Cycle Adjustment Mechanism

Assume that let N denote the number of potential neighbor nodes in the boundary nodes assist scheduling process. The nodes within the region S are neighbor nodes which are not found by Dis. According to the characteristics of Poisson distribution, we know that the expected number of nodes in $S^{'}$ are as follows.

$$E[Sum^{S'}] = \lambda \Delta S' \tag{1}$$

Where $\Delta S^{'}$ is the area of the S region.

4.1 Calculate $\Delta S^{'}$

The distance between the centers of two communication circles is *distD*. Shown in Fig.2, the intersection area of two communication circles is ΔS, the areas of the triangle TAB and TAB are the same, which is represented by ΔS_{TAB} and $\Delta S_{T'AB}$, respectively.

$$\Delta S = 2 * (\alpha R^2 - \Delta TAB) \tag{2}$$

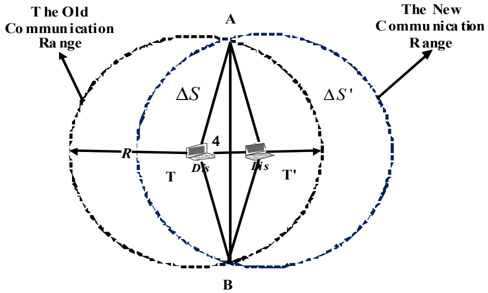


Fig. 2. The area diagram

Where $\Delta TAB = \frac{1}{2} dist_D R \sin\alpha$, $\alpha = \arccos\frac{dist_D}{R}$. The area can be obtained by formula (1) and formula (2).

$$\Delta S' = \pi R^2 - \Delta S \tag{3}$$

4.2 Duty Cycle Adjustment Process

According to the formula (3), it can be obtained the expected number of nodes $E[Sum^{S'}]$ in S.

$$E[Sum^{S'}_{no}] = E[Sum^{S'}] - N = \lambda\Delta S' - N \tag{4}$$

Assumed that the wake-up time of *Dis* is W_{Dis} and the wake-up cycle is T, we can take the wake-up time segment W_{Dis} as box and $E[Sum^{S'}_{no}]$ nodes as ball, then based on the model of balls and boxes, we know that p_0^{wake} that the wake-up time of the $E[Sum^{S'}_{no}]$ nodes which are not found in S is not within the wake-up time of *Dis*.

$$p_0^{wake} = \left(1 - \frac{W_{Dis}}{T}\right)^{E[Sum^{S'}_{no}]} \tag{5}$$

In the formula (5),W_{Dis}/T indicates the probability that single node is woken up within the wake-up time of *Dis*. p^{wake} that there are some of $E[Sum^{S'}_{no}]$ nodes which are woken up within the wake-up time of the node *Dis* is as follows.

$$p^{wake} = 1 - p_0^{wake} \tag{6}$$

However, some nodes are not woken up in W_{Dis} , and then it needs to adjust the duty cycle of *Dis* to discover the neighbor nodes. According to the obtained probability p_0^{wake} and the wake-up time W_{Dis} , the wake-up time extension is $W_{Dis} * p_0^{wake}$, and the duty cycle $duty_{dis}$ is adjusted as follows.

$$duty_{dis} = \frac{W_{Dis} + W_{Dis} * p_0^{wake}}{T} \tag{7}$$

4.3 F-Node Scheduling

The variation of the wake-up time cycle is k ($0 < \mathrm{k} \leq 1 + T/2W^{'}_{Dis}$). Let L_i define the node's position of the current wake-up time in the cycle T, and L_0 =0 denote that the wake-up time is at the foremost end of wake-up cycle T, then we know L_{i+1} is as follows.

$$L_{i+1} = ((L_i * W^{'}_{Dis} + W^{'}_{Dis})/W^{'}_{Dis}) mod k \tag{8}$$

We can get L_1=1,..., L_{k-1}=k-1 by the formula (8), which denotes the node's position of the current wake-up time in the cycle T, respectively. Shown in Fig.3, L_1 is at the second time period in a cycle, and then we can know the position indicated by L_{i+1}. Algorithm 1, called NBRD algorithm, is as follows.

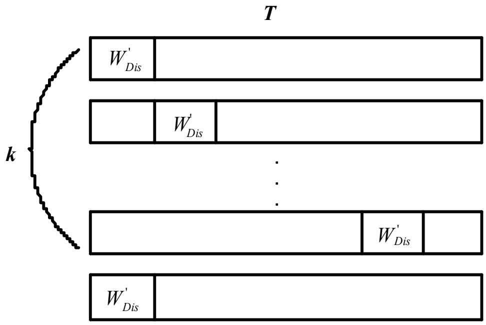


Fig. 3. F-node Scheduling

Algorithm 1. NBRD algorithm.

```
Input: $\lambda$ , F-node speed $v$,other node speed $v$;
Output: the number of discovered neighbors;
Obtain the boundary set of $top - \theta$.
Allocate the wake-up time of the boundary nodes in $top - \theta$:$W_1$, $W_2$ ,$W_3$ ,..., $W_{\omega_1}$, set $W_1 \cap W_2 \cap W_3 \cap \ldots \cap W_{\omega_1} = \Phi$ .
Calculate $\Delta S'$ and predict the number of nodes in the communication radius.
Utilize the model of the ball and the box to calculate the wake-up time and adjust the duty cycle.
Set the variation of the wake-up time cycle $k$.
Apply the formula $L_{i+1} = ((L_i * W_{Dis}^{'} + W_{Dis}^{'})/W_{Dis}^{'}) \bmod k$ to calculate the location of the wake-up time in the cycle.
Schedule the wake-up time as a cycle with $k$, until finish the monitor of $k$ cycles.
```

5 Theoretical Analysis

Due to the nodes following Poisson distribution in the networks, we apply poisson point process [9] to analyze the regulatory mechanisms of the F-node.

5.1 The Probability of New Discovered Neighbors in S

According to the nature of poisson point process, the relationship between the area and the probability that there are no nodes in a certain area is as follows.

$$v_K = \exp(-\lambda V(K)) \tag{9}$$

Where $V(K)$ is the area of a region. The probability that there are no nodes in S is as follows.

$$v_{S'} = \exp(-\lambda \Delta S') = e^{-\lambda \Delta S'} \tag{10}$$

So, we can obtain that there is a node in S.

$$v_{S'}^{0+} = 1 - v_{S'} = 1 - e^{-\lambda \Delta S'} \tag{11}$$

From the formula (11), we know that the area of S is greater, the probability that there is a node in S is greater, so the adjustment of the duty cycle is more meaningful.

The expected number of the undiscovered nodes in S are $\lambda \Delta S' - N$, and the expected number of all the nodes in S are $E[Sum^{S'}]$ and the probability that there are new nodes in S is $p_n = (\lambda \Delta S' - N) / E[Sum^{S'}]$. Based on the nature of poisson point process, we can get $v_{S'}^{0+}$, so p_d that it can discover new nodes in S is $p_d = pE[Sum_{no}^{S'}]$ is not zero |there exist some nodes in the region S= $v_{S'}^{0+} * p_n = \left(1 - e^{-\lambda \Delta S'}\right) * \frac{\lambda \Delta S' - N}{E[Sum^{S'}]}$.

5.2 Successfully Monitor Rate

It is assumed that the probability of successfully monitoring new nodes in k cycle is p_s, and the probability of failing to monitor the nodes waking up is p_s, so p_s+p_s=1. p_s can be launched by the birthday paradox.

$$p_s{}' = \prod_{i=1}^{k} (1 - \frac{iW'_{Dis}}{T}) \tag{12}$$

Based on p_s+p_s=1, we can obtain p_s:

$$p_s = 1 - \prod_{i=1}^{T/2W'_{Dis}+1} (1 - \frac{iW'_{Dis}}{T}) \tag{13}$$

When $k = T/2W'_{Dis} + 1$, p_s reaches the maximum value, so in the wake-up time of Dis, the probability of successfully monitoring neighbor nodes is at most $1 - \prod_{i=1}^{T/2W'_{Dis}+1} (1 - \frac{iW'_{Dis}}{T})$.

5.3 Discovery Delay

In the process of moving, a node applies the duty cycle to find its neighbor nodes. When the movement direction changes after time t, Dis discovers neighbor nodes, and let delay denote the discovery delay of Dis, so delay is as follows.

$$delay = T_d + T_c + T_{schedule} \tag{14}$$

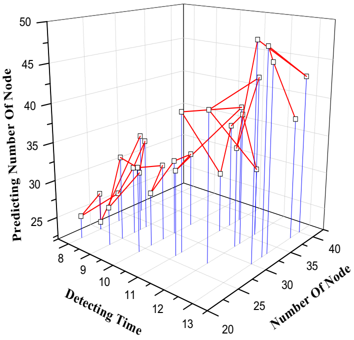
Where T_d is the required time when Dis applies boundary node to discover potential neighbors, T_c is a wake-up cycle of Dis, $T_{schedule}$ is the time consumed to periodic schedule the wake-up time, $T_{schedule} = k * T$, so we get $delay = (k + 2) * T + T(\omega_1 - 1)/\omega_2$.

6 Experimental Results

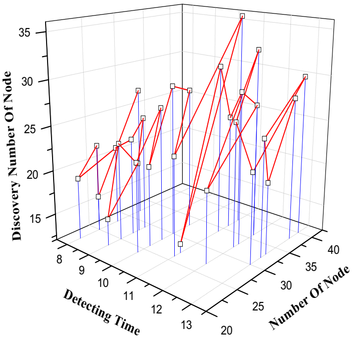
This section verifies the performance of the NBRD algorithm via simulation experiments. Be sake of convenience, *top*-1 and *top*-2 stand for the boundary nodes in the set of $top - \theta$ and $top - 2\theta$, respectively.

6.1 The Impact on the Number of Discovered Neighbors for the Monitoring Time

In Fig.4(a), it is more accurate prediction of the number of nodes in the majority of cases, and there are larger differences between the predicted number of nodes and the actual number of nodes in the rare cases. When the prediction is not accurate, the number of discovered neighbors is less which can be seen from Fig.4(b), because the predicted number of nodes has an impact on the adjustment of the duty cycle of *Dis*, which makes the adjustment of the duty cycle of *Dis* not accurate.From Fig.5, we know that the results are better for the NBRD algorithm using the nodes in the set of $top - 2\theta$ to the prediction, because the number of nodes within the set of $top - 2\theta$ is more, so that the more nodes are woken up in one wake-up cycle to assist *Dis* to discover more potential neighbors.



(a)



(b)

Fig. 4. The monitoring time VS. the number of discovered nodes

6.2 The Impact on the Number of Discovered Nodes for Delay

Shown in Fig.6, when the given delay is closed to 4000 unit time and 6000 unit time, the number of neighbor nodes discovered by NBRD algorithm using the set of $top - \theta$ is larger than that using the set of $top - 2\theta$. The reason is that the wake-up time of nodes is more concentrated, which makes the number of discovered nodes more in this two cases. While in the other cases of delay, the number of discovered nodes using $top - 2\theta$ is more. This is because the number of nodes in $top - 2\theta$ is more, which can help Dis better to discover its neighbor nodes, and with the more nodes helping discover potential neighbor nodes within the communication range, it can properly adjust the duty cycle of the node Dis, which makes Dis be woken up in time and discover more neighbor nodes.

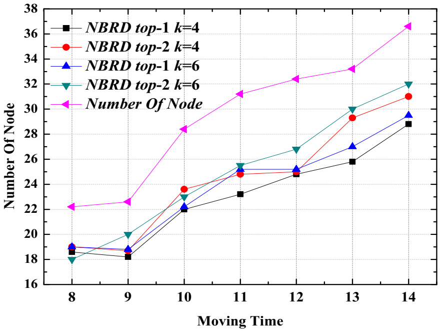


Fig. 5. The monitoring time VS. the number of discovered nodes

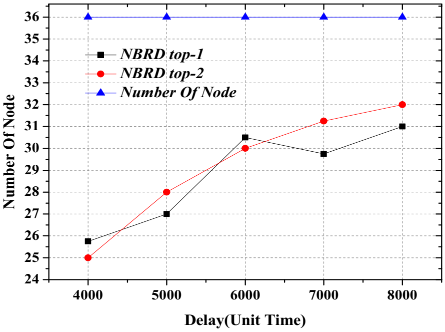


Fig. 6. Given delay VS. the number of discovered nodes

6.3 The Impact on Energy for the Monitoring Time

In Fig.7, with the increase of the initial monitoring time, its energy consumption is gradually reduced, and finally reaches a steady state. The reasons are that the increase of the initial monitoring time leads to the gradual increase of the new communication region and the number of nodes in the region. When the initial monitoring time is smaller, the number of nodes is less and the predicted number of potential nodes is also less, and then the dynamical adjustment of the duty cycle becomes larger after a probability calculation. After the monitoring time increases, the predicted number of potential nodes become lager and then tend to stabilize, making the adjustment of the duty cycle and the energy consumption in the process of neighbor discovery stable.

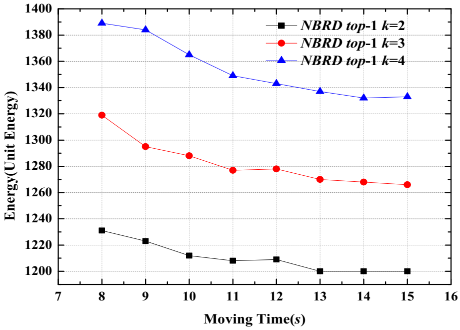


Fig. 7. The monitoring time VS. energy

6.4 The Impact on Energy for the Scheduling Number *k* of the Wake-Up Time

Alpha represents the angle in the figure. Shown in Fig.8, the energy consumption is increasing with the scheduling number of the wake-up time increasing, The reason is that the wake-up times are increasing so that the energy consumption increases, when the scheduling number of the wake-up time increases. When using the boundary nodes in $top - 2\theta$, compared with using $top - \theta$, the energy consumption has a slight increase at $k = 6$, but at the other k, its energy consumption is almost flat. Finally, we know that the selected boundary nodes have no effect on energy consumption, and the value of k has a greater impact on energy consumption.

6.5 The Impact on the Number of Discovered Neighbors for Network Topology

Shown in Fig.9, in the R-axis, with R changing, the number of nodes that exists in the communication region are increasing, and the number of discovered neighbor nodes is also increasing with the value of R changing. We can see the

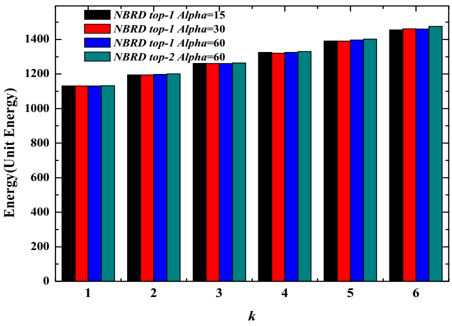


Fig. 8. The scheduling time *k* VS. energy

gap between the number of nodes in the communication region and the number of the discovery nodes is within a certain range in most cases, and in rare cases, the number of nodes in the gap have a larger difference, so the effect of discovering nodes is better. For the individual conditions, NBRD algorithm can also obtain good experimental results, the reasons are that the duty cycle is adjusted by estimating the number of nodes predicted in the new region, changing the wake-up time of the node *Dis*, so that the node are woken up in time to monitor the neighbor nodes, and the scheduling cycle of the wake-up time *k* selected is 4, covering a greater range of the cycle.

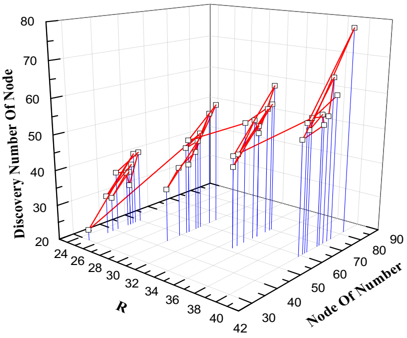
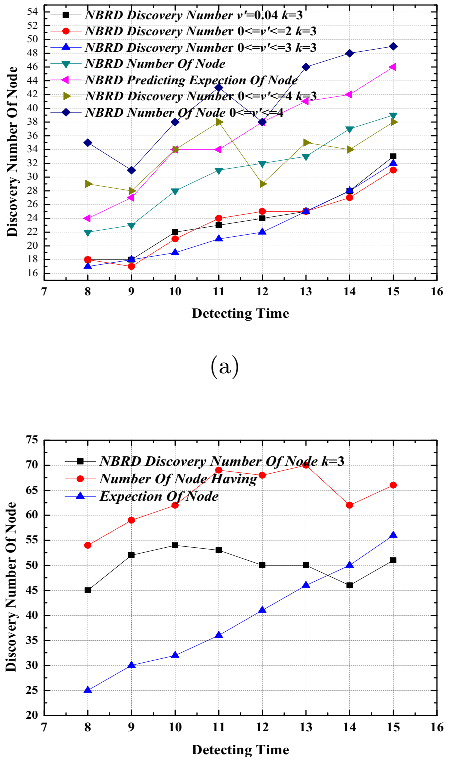


Fig. 9. Network topology VS. the number of discovered nodes

6.6 The Impact on the Number of Discovered Nodes for the Moving Speed of Nodes

The impact on the number of discovered neighbor nodes for the different moving speed of the node is shown in Fig.10 (a). The experimental results using the speeds of *Dis* which are randomly selected from the range (0,2), (0,3), (0,4) and



(a)

(b)

Fig. 10. Moving speed VS. the number of discovered nodes

0.04m/s are very similar, the difference between the predicted results and the actual number of nodes is small, and the experimental results using 0.04 m/s are better in most cases. The reasons are that the speed is small and the network change has certain stability, and the results are predicted more accurately and the adjustment of the duty cycle is more precise. When the speed of all nodes is 4 m/s, the results are shown in Fig.10 (b). When all the predicted results are increasing with the initial monitoring time increasing, the difference between the predicted results and the actual number of nodes is larger, and the reasons are that the speed 4 m/s is the same with the speed of *Dis*, and the network topology more rapidly changes with the rapid movement of the nodes, but the results are predicted based on poisson point process, so the rapid changes of the network topology make the distribution of nodes not satisfy the characteristics of Poisson point process, therefore the difference between the predicted results and the actual situation is larger.

7 Conclusion

Aiming at the problem of neighbor discovery , this paper provides an algorithm based on dynamically adjusting the duty cycle. This algorithm firstly predicts the number of potential neighbors and estimates the number of neighbor nodes in the new region, then uses the model of ball and box to calculate the duty cycle of the nodes and real-time adjusts the duty cycle in order to wake up the F-node in time. Finally, after adjusting the duty cycle, we can periodically schedule the wake-up time in the wake-up cycle to monitor neighbors, and increase the coverage to complete the monitoring of neighbor nodes. Simulation results show that in the process of neighbor discovery, the nodes not only use less energy, but also discover more neighbor nodes in a shorter time.

Acknowledgments. This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant Nos.61070193, Heilongjiang Province Founds for Distinguished Young Scientists under Grant Nos.JC201104, the Foundation of University Science and Technology Innovation Team Building Program of Heilongjiang Province under 2011PYTD002, Harbin Municipal Science and Technology Innovation Talent research special funds outstanding academic leaders funded project under Grant Nos. 2011RFXXG014, Heilongjiang Province Science and Technology Innovation team building program under Grant Nos. 2011PYTD002, Program for New Century Excellent Talents in subsidy under Grant Nos. NCET-11-0955, the Heilongjiang Provincial Department of Education for New Century Excellent Talents Support Program funded under Grant Nos. 1252-NCET-011.

References

1. McGlynn, M.J., Borbash, S.A.C.: Birthday protocols for low energy deployment and flexible neighbor discovery in ad hoc wireless networks. In: ACM MOBIHOC, pp. 137–145. ACM Press, New York (2001)
2. Dutta, P., Culler, D.C.: Practical asynchronous neighbor discovery and rendezvous for mobile sensing applications. In: SenSys, pp. 71–84. IEEE Press, New York (2008)
3. Kandhalu, A., Lakshmanan, K., Rajkumar, R.R.C.: U-connect: a low latency energy-efficient asynchronous neighbor discovery protocol. In: IPSN, pp. 350–361. IEEE Press, New York (2010)
4. Tseng, Y.C., Hsu, C.S., Hsieh, T.Y.C.: Power-saving protocols for ieee 802.11-based multi-hop ad hoc networks. In: INFOCOM, pp. 200–209. IEEE Press, New York (2002)
5. Lai, S., Ravindran, B., Cho, H.C.: Heterogenous Quorum-Based Wake-Up Scheduling in Wireless Sensor Networks. IEEE Transactions on Computers, 1562–1575 (2010)
6. Karowski, N., Viana, A.C., Wolisz, A.C.: Optimized Asynchronous Multi-channel Neighbor. In: INFOCOM, pp. 536–540. IEEE Press, New York (2011)

7. Mehedi, B., Robin, K.C.: SearchLight: A Systematic Probing-based Asynchronous Neighbor Discovery Protocol. In: ACM SIGMOBILE, pp. 71–76. ACM Press, New York (2010)
8. Vasudevan, S., Towsley, D., Dennis, D., Khalili, R.C.: Neighbor discovery in wireless networks and the coupon collectors problem. In: ACM MobiCom. ACM Press, New York (2009)
9. C.: Statistical Analysis and Modeling of Spatial Point Patterns
10. Purohit, A., Priyantha, B., Liu, J.C.: WiFlock: Collaborative group discovery and maintenance in mobile sensor networks. In: IPSN, pp. 37–48. IEEE Press, New York (2011)
11. Karowski, N., Viana, A.C., Wolisz, A.C.: Optimized Asynchronous Multi-channel Neighbor Discovery. In: INFOCOM, pp. 536–540. IEEE Press, New York (2011)
12. Zhang, P., Sadler, C.M., Lyon, S.A., Martonosi, M.C.: Hardware design experiences in zebranet. In: SenSys., pp. 227–238. IEEE Press, New York (2004)

Diversity between Human Behaviors and Metadata Analysis: A Measurement of Mobile App Recommendation

Xiao Xia[*], Xiaodong Wang, and Xingming Zhou

School of Computer Science, National University of Defense Technology,
Changsha, P.R. China, 410073

Abstract. The explosive growth of mobile apps has given rise to the significant challenge of app discovery. To meet this challenge, the Google Play market utilizes the user behaviors data to provide app recommendations. By making use of experiences of the user crowd, such recommendations are of help to users for discovering apps. However, they are concurrently restricted to the local scope of the user experiences, as most users have only accessed a limited amount of apps. To conquer this constraint, we propose a novel recommending method by utilizing the global information of apps. To be specific, we leverage the Latent Semantic Indexing method to analyze the metadata of apps, which is globally held by the market. We thus obtain the similarity measurements among apps and based on them we generate app recommendations. To further understand both the human behavior based and the metadata analysis based methods, we then measure the diversity within them from multiple levels and scopes. Through such measurements, we eventually discover new knowledge of user preferences and gain better understanding of both recommending methods. These observations further indicate that there are necessities and potentials to evolve the existing mobile app recommender systems by integrating new recommending methods.

Keywords: Mobile app recommendation, human behavior, metadata analysis, diversity measurement.

1 Introduction

Recent years have witnessed an explosive growth in the population of mobile apps. As reported by the latest investigations, both the Apple Store and the Google Play have hit the milestone of 700,000 apps in their markets [1]. The number of mobile apps is increasing so quickly that it turns out a significant challenge for users to find out apps of interest among the huge amount. To facilitate this problem, recommender systems are utilized by online app markets to provide users with app suggestions. For instance, the Google Play market

[*] This work is supported by the projects of National Natural Science Foundation of China: *No. 61070201*, *No. 61170260* and *No. 61202486*.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 300–312, 2013.

recommends apps that "users viewed this app also viewed" and "users installed this app also installed". Obviously, such recommendations are generated based on the user behaviors, which are assumed to leverage some kind of Collaborative Filtering (CF) approaches. These recommendations help users discover new apps by making use of experiences of the user crowd. However, solely adopting the CF approach would also retain its intrinsic limitations. Specifically, recommendations can be restricted by the experiences of users, most of whom only know a limited number of apps.

To eliminate the above constraint, we propose a novel method for recommending mobile apps by making use of the globally hold information, i.e., the metadata of apps. We leverage the method of *Latent Semantic Indexing* to analyze the metadata of apps. Through the mining of latent semantics laying in the metadata, we obtain similarity measurements among apps. We then generate app recommendations based on such measurements, which adopts a Content Based (CB) approach. Diverse from the CF approach, it makes better use of the global information of apps whose amount is much larger than that of the apps each user have accessed.

To further understand the CF and CB methods in the context of mobile app recommendation, we then measure the diversity within them based on the real data of 103348 apps. The measurements are conducted from a variety of levels, i.e., the node, the set and the network level. Through such measurements, we eventually discover new knowledge of user preferences and derive better understanding on the features of each method. By these means, we concurrently identify the necessity of evolving the existing system and indicate the potential of integrating two approaches for the evolution.

Main efforts and contributions of this paper are listed below:

1. We propose a novel method for recommending mobile apps by making use of the global app information to overcome the constraint of user experiences. To be specific, we leverage the LSI method to analyze the metadata and measure the similarity of apps, based on which we generate app recommendations.
2. We present original diversity measurements between the CF and CB methods in the domain of mobile app recommendations, using a large scale of real data. Through the measurements, we discover new knowledge of user preferences and derive better understanding on the pros. and cons. of both approaches.

Through above efforts, we improve the understanding of mobile users and app markets and identify the necessity and potential of evolving the mobile app recommendations. By this way we call for more efforts dedicated to the evolution of mobile app recommender systems. In the reminder of the paper, we firstly describe the data set we build for the measurements in Section 2. We then propose our recommending method and describe two kinds of app recommendations in Section 3. Afterwards, we present measurements of the diversities within two methods in Section 4. Finally we draw conclusions in Section 5.

2 Data Set

In this section we describe the data set which is built to measure the recommendations provided by both the Google Play market and our method. These two recommendations are separately generated based on the *human behaviors* using a CF approach and based on the *metadata analysis* using a CB approach. For simplicity, we refer them as the HB and MA recommendations, respectively.

To describe the HB and MA recommendations, we crawl web pages from the Google Play market to build our data set. To reflect the real app usage of users, we start the data collecting from the latest apps, the most popular apps and randomly selected apps. Therefore, we cover the start points from which most users begin their surfing in the app market. We stop when there are no new apps found following the links on the websites thus the data set is assumed to cover apps that most users would access at most of the time.

After the crawling, we parse the web pages to extract information for our study. For the measurement of HB recommendations, we extract the *identifiers* and the *recommending links* of all the apps. Recommending links are web links which point from the web page of an app to the web pages of its HB recommendations. For the description of MA recommendations, we gather the metadata of apps which mainly includes the *category* and the *description.* To characterize the diversities within two recommendations, we further collect other information such as the *prices* and *installations* of apps. After all, we have built a data set of 103348 apps within all the 29 app categories.

3 Tale of Two Recommendations

In this section we propose our recommending method and describe the two kinds of app recommendations, which are generated separately by the Google Play market and our method.

3.1 Recommendation Based on Human Behaviors

Google Play provides mobile users with app recommendations that "users viewed this app also viewed" and "users installed this app also installed", which are referred as "alsoview" and "alsoinstall" apps for simplicity. Obviously, they are generated based on the behaviors of mobile users.

To measure this kind of recommendations, we construct the "behavior" network G_b to capture both the apps and recommending links among them. Additionally, although the "alsoview" and the "alsoinstall" apps collaboratively serve as HB recommendations, they may perform differently on doing so. Therefore, we also extract the "alsoview" network G_v and the "alsoinstall" network G_i. In these networks, the nodes represent the apps and the edges denote the recommending links. For instance, e_{kj} in G_v denotes that app j is recommended for k as an "alsoview" app.

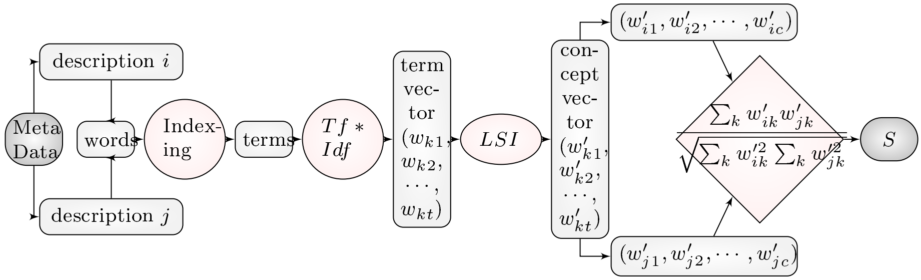

Meta Data
description i
description j
words
Indexing
terms
$Tf * Idf$
term vector $(w_{k1}, w_{k2}, \cdots, w_{kt})$
LSI
concept vector $(w'_{k1}, w'_{k2}, \cdots, w'_{kt})$
$(w'_{i1}, w'_{i2}, \cdots, w'_{ic})$
$(w'_{j1}, w'_{j2}, \cdots, w'_{jc})$
$\frac{\sum_k w'_{ik} w'_{jk}}{\sqrt{\sum_k w'^2_{ik} \sum_k w'^2_{jk}}}$
S

Fig. 1. Process of metadata similarity measurement using LSI

3.2 Recommendation Based on Metadata Analysis

Since most of users have not accessed a large number of apps, the HB recommendations can be restricted by the user experiences. To eliminate this constraint, we turn to make better use of the global information of apps. We mainly focus on the *app descriptions* since they are not only the primary channels for developers to specify their apps, but also the main references for users to understand the apps. Thus we assume that similar descriptions would suggest similar apps. We then recommend similar apps to users by measuring the similarity among metadata to capture the similarity among apps.

To know better about the semantics of descriptions and derive reliable measurements, we introduce the method of LSI (Latent Semantic Indexing), whose effectiveness has been verified in a variety of IR tasks [2]. LSI uses the *Vector Space Model* to represent documents by vectors of weighting terms. It then projects the term-document matrix to a lower-dimensional space. By reducing the dimensionality, LSI mines the meanings and the variability of terms underlying the documents, i.e., the synonymy and polysemy. After all, the original term-document space is projected to the semantic space, which represents the semantic concepts instead of raw terms. Then the similarity measurement is expected to gain a better understanding when it is based on the concepts comparison.

As illustrated in Figure 1, we obtain the app similarity by measuring the description similarity, based on which we provide the most similar apps as our MA recommendations. To capture the recommending relationships of MA recommendations, we also construct the recommending network G_m. Similar to G_b, nodes of G_m represent the apps and an edge e_{ij} in G_m denotes that we recommend app j for app i. According to the general number of HB recommendations, we provide the top 8 similar apps as the MA recommendations.

4 Diversity Measurement

In this section, we measure the diversities within the HB recommendations and the MA recommendations from various levels, i.e., the app level, the set level and

Table 1. The category and price assortativity of recommending networks

networks	category assortativity	price assortativity
G_v	0.659	0.244
G_i	0.691	0.773
G_b	0.675	0.503
G_m	0.415	0.198

the network level. Through the measurements, we are to discover new knowledge of user preferences and motivations by implicit elicitation [3] based on the data analysis. Meanwhile we obtain better understanding on the pros. and cons. of the CF and CB approaches in recommending mobile apps. Both kinds of investigations cooperatively identify the necessity and potential of evolving the existing industrial-strength recommender system.

4.1 Diversity on App Level

On the app level, we measure the diversity between HB and MA recommendations with respect to the app properties. Especially, we investigate whether the recommended apps for one app are of homogeneous to it from the view of app properties. To be detailed, we focus on the category and the price properties in this paper. Therefore, the measurements capture the questions below:

- Whether and to what extent the app recommendations for i are in the same category with i;
- Whether and to what extent the app recommendations for a free/paid app are free/paid.

These measurements examine the capability of recommendations on capturing the needs of users and the expectations of markets. Specifically, recommending more apps that are in different categories would help users to discover novel apps, while recommending free or paid apps influences the revenue of the online market.

To conduct the measurements, we introduce the metric *attribute assortativity* from the complex network analysis. Attribute assortativity in a network identifies the preference of nodes to connect to those with similar properties. Take the recommending network for instance, if free apps prefer to connect to free apps rather than paid ones, there is a high price assortativity in the network. We adopt the definition of assortativity coefficient proposed in [4] as:

$$Aa = (Tr\mathcal{E} - |\mathcal{E}|^2)/(1 - |\mathcal{E}|^2), \qquad (1)$$

where $\mathcal{E}$ is the matrix consisted of elements e_{ij}, which denotes the fraction of edges connecting a node with property i to another node with property j.

We examine the “category” and “price” assortativity of the recommending networks constructed in Section 3. From the results listed in Table 1, we can see that there is a higher category assortativity in the HB recommendations, which indicates that *from one app users prefer to view and install more apps that are of the same category.* Recommendations based only on such preferences may narrow the scope of app discovery thus to prevent users from finding out more novel apps. Reversely, the MA recommendations which utilize the global information of apps exhibit lower category diversity, thus can recommend apps from more categories to users.

Looking into the price assortativity, there is a balance (0.503) in the overall preferences of users to find free or paid apps. However, a significant diversity shows up within diverse behaviors, i.e., the "alsoinstall" apps are much more homogeneous than the "alsoview" apps. It suggests that *users who have installed free/paid apps prefer to install more same ones while users who have viewed free/paid apps tend to view more different ones.* Motivations behind such behaviors may lay in that users tend to surf freely across web pages while they are more cautious on consumption (installing). Such observations can play an important role on evolving the recommender system when taking the profit expectations of online markets into consideration. The low price assortativity of G_m suggests that *for a paid/free app, there are more free/paid apps that are similar to it.* This observation may suggest the motivations of the developers who choose to distinguish their apps from the similar ones by their pricing strategies.

4.2 Diversity on Set Level

On the set level, we investigate the diversity within the recommendation sets provided by different methods. To be detailed, given an app i, we denote sets $R_b(i)$ and $R_m(i)$ as apps recommended by the HB and MA methods, respectively. We study the diversity between the two sets thus to identify differences between two methods while studying motivations of users and limitations of the CF method. On this level, we measure diversities from the view of coverage and ranking.

Diversity by Coverage. Firstly we examine that whether recommendations provided by one method are also recommended by the other. To capture this feature, we characterize the $R_b(i)$ and $R_m(i)$ on their coverage over each other. This diversity is partly for identifying the differences between the two thus to indicate the potential of their collaboration. On the other hand, we take a look into the preferences of user behaviors and the features of the online market. Specifically, the measurements capture the questions:

- *Whether recommendations provided by two methods diverse significantly?*
- *Whether mobile users prefer viewing or installing similar apps?*
- *How many similar apps have been covered by the online market?*

We introduce the metric *precision* to characterize the coverage between sets:

$$Pr_{R_1->R_2} = |R_1 \cap R_2|/|R_1|, \tag{2}$$

Table 2. Diversity measurement settings

Parameter	$Pr(i)_{R_b^{nhop}->R_m^{ntop}}$	$Pr(i)_{R_m^{ntop}->R_b^{nhop}}$
nhop	1-4	1-4
ntop	100	1-8

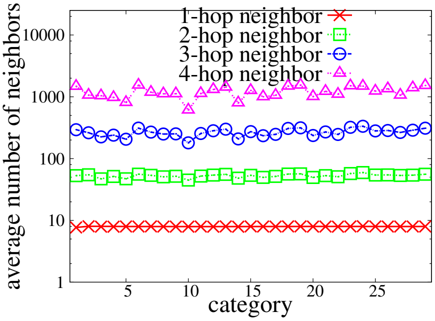
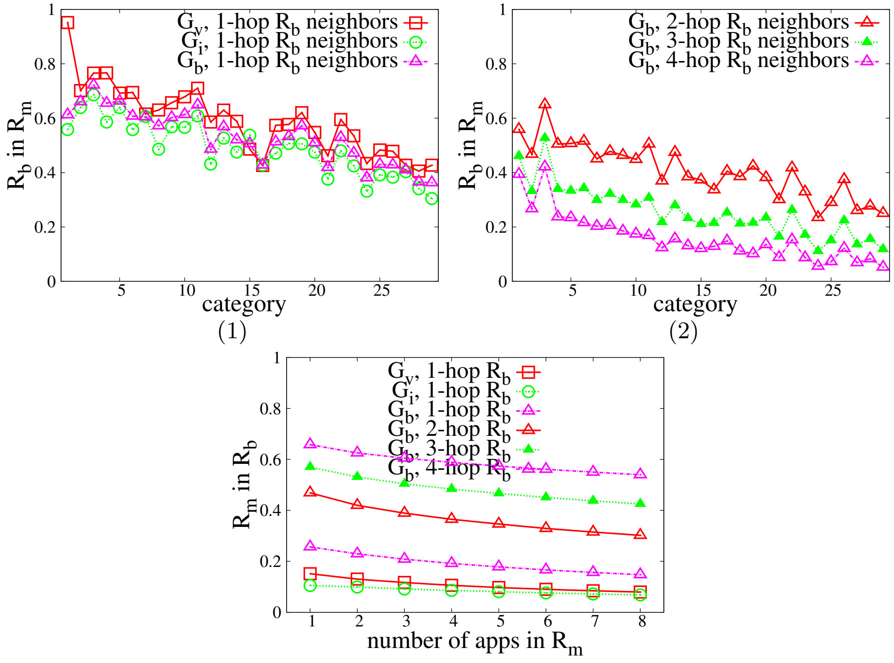


Fig. 2. Number of n-hop neighbors recommended based on human behaviors in the market

where R_1 and R_2 are different sets of apps. Using this metric, for each app i, we characterize the precisions of recommendations from both directions. That is, we measure both the $Pr(i)_{R_b^{nhop}->R_m^{ntop}}$ and $Pr(i)_{R_m^{ntop}->R_b^{nhop}}$. The *ntop* denotes the number of recommended apps because our MA method provides *TopN* recommendations. The *nhop* define the hops between an app and the recommended ones for it on the web pages, as apps recommended by the HB method are presented in the form of links on the web pages.

The measurement settings are listed in Table 2. We denote *nhop* to 1-4 as Figure 2 shows that the number of n-hop neighbors in G_b grows exponentially. Thus we assume that 4-hop neighbors are sufficient to capture the feature of HB recommendations. Note that for n-hop neighbors when $n > 1$, we only measure the overall HB recommending network G_b. The reason is that measurements of G_v and G_i would make little sense, as users would rarely follow only one kind of recommending links all along. We set *ntop* to 100 when measuring the $Pr(i)_{R_b^{nhop}->R_m^{ntop}}$, because we care about how many HB apps are in the top similar list of MA recommendations, while 100 is a usual length of top rankings with respect to the number of app neighbors . We set *ntop* to 1-8 when measuring the $Pr(i)_{R_m^{ntop}->R_b^{nhop}}$, as Google Play recommend 8 HB apps for most apps, so that we investigate the precision of MA method by recommending the same amount of apps. Note that in the measurements we only investigate the coverage between apps of the same categories. This is partly because users tend to view or install apps of same categories, as observed in Section 4.1. Moreover, the similarity measurements of the same category are assumed to be more confident, because apps of same category that adopt similar descriptions can be assumed to be "really" similar. This to some extent avoids possible errors brought by LSI.



(1)

(2)

(3)

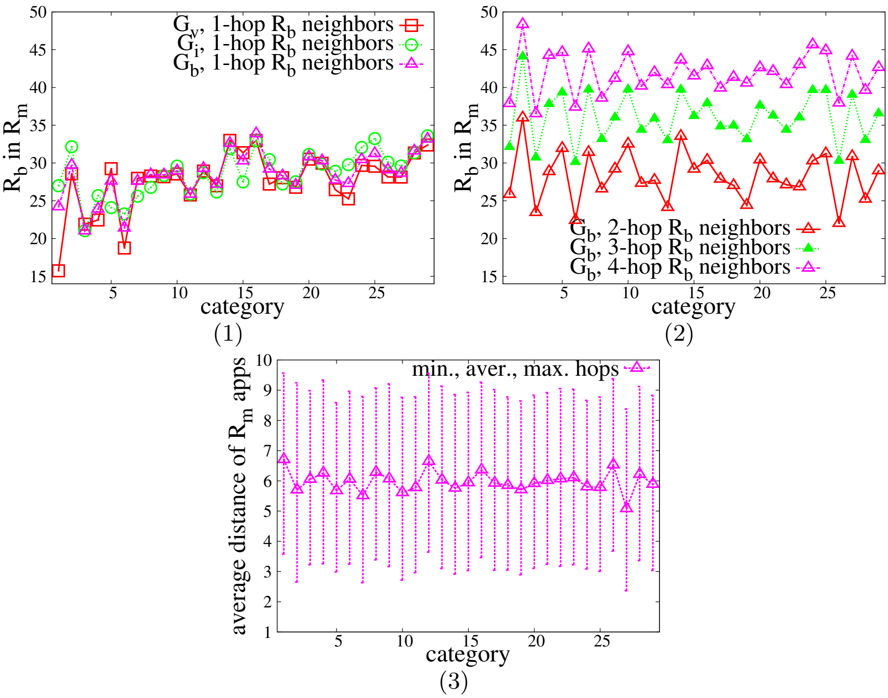
Fig. 3. Diversity measurement by coverage, HB recommendations in MA ones(1,2) and MA recommendations in HB ones(3)

Results of diversity measurement by coverage are illustrated in Figure 3. We can see that in most cases each of the two recommendations achieves a low precision ($< 50\%$) to cover the other, thus the HB and MA recommendations are diverse on recommending apps to users. Moreover, from Figure 3:(1,2) we see that *there are indeed similar apps found by human behaviors*, especially in the 1-hop R_b neighbors which users can directly access on web page of one app. This indicates that the mobile users are actually finding similar apps. However, Figure 3:3 shows that *only a part of the top similar apps are covered by the existing system.*

Diversity of Ranking. Then we investigate the diversity between the HB and MA recommendations based on the app rankings. Driven by the similar motivations, this diversity serves as a further step of the diversity of coverage. It characterizes quantitative indicators for deep understanding of the user preferences and the system features. Specifically, the measurements are to capture the questions below.

– *How similar are the apps that have been found by users?*
– *How many hops would it take for users to reach the most similar apps?*

To measure such a diversity, we define the metrics average ranking Ar and average distance Ad for each app i.



(1)

(2)

(3)

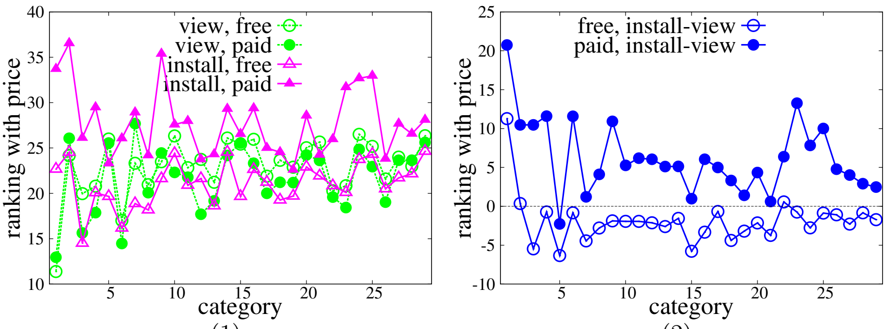
Fig. 4. Average ranking of HB recommendations in metadata analysis(1,2) and Average distance of MA recommendations in online market(3)

$$Ar(i) = Average\{r_x \mid x \in R_b(i) \wedge x \in R_m(i)^{r_x} \wedge x \notin R_m(i)^{r_x-1}\}, \qquad (3)$$

$$Ad(i) \; = Average\{d_x \mid x \in R_m(i) \wedge x \in R_b(i)^{d_x} \wedge x \notin R_b(i)^{d_x-1}\}. \qquad (4)$$

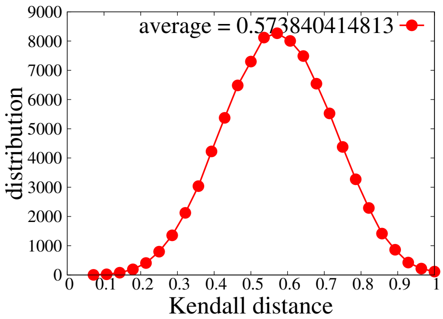
We use Ar to measure the rankings of HB recommendations with respect to the LSI similarity measurements. This is to examine how similar are the apps that have been found by users. Meanwhile, through the Ad we measure the distances of MA recommendations on the web pages of the market. This is to examine how many hops it needs for users to reach the most similar apps in the existing market.

Similarly to the diversity measurement by coverage, we investigate the Ar of HB recommendations which are in the top-100 list and recommend top 1-8 MA recommendations when measuring the Ad. From the Figure 4:(1,2) we see that *the average ranking of apps recommended by the human behaviors holds a small value, which however is not very small at the same time.* This indicates that on one side users tend to find out similar apps, while on the other side they fail to discover the most similar ones. We attribute this to the HB method of the existing system, as it makes use of users' judgments while concurrently retaining the constraints of user experiences. Further verified by Figure 4:3, *it may take many jumps to reach the most similar apps crossing web pages.*



(1) (2)

Fig. 5. Average ranking of alsoview and alsoinstall apps(1); Ranking distance between alsoinstall and alsoview apps, i.e., $A_r^{alsoinstall} - A_r^{alsoview}$(2)



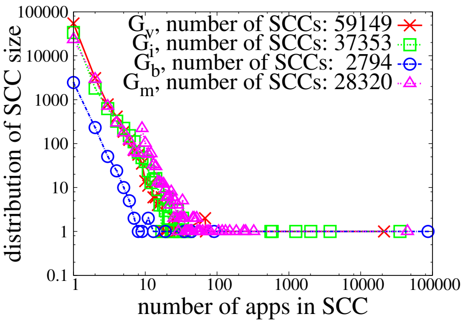


Fig. 6. Distribution of normalized Kendall distances

Fig. 7. Distribution of SCC sizes in networks

To better understand the behaviors of users, we take a further look into the ranking diversity within the "alsoview" and "alsoinstall" recommendations. We measure the ranking diversity of the "alsoview" and "alsoinstall" apps separately in the categories of free apps and paid ones. As shown in Figure 5, *users tend to install apps of more similar for free apps while they prefer to install apps of less similar for paid apps.* Motivations of such observations may lie in that users tend to try more free similar apps while they prefer buying more apps that are dissimilar. This suggests the system to take the app price into account when choosing to recommend similar or diverse apps to users.

Both the Ar and Ad investigate a single ranking. We further introduce the Kendall's tau distance to measure the diversity within two rankings. Specifically, for a set of apps, we measure the distance between the rankings which are separately generated by the HB and MA methods. To this end, we derive app sets from the MA recommendations $R_m(i)$, as the HB recommendations do not hold ranking information. Then their ranking in the MA recommendations, denoted

Table 3. Diversity measurement on the network level

Network	Link reciprocity	Clustering coefficient
G_v	0.151	0.217
G_i	0.323	0.239
G_b	0.308	0.192
G_m	0.389	0.315

as τ_i^m, is sorted by their similarities with app i. The ranking in the HB recommendations, denoted as τ_i^b, is sorted by their distances from i. Then the Kendall distance is defined as:

$$K(i) = |\{(j,k) : j < k, \tau_i^m(j) < \tau_i^m(k) \wedge \tau_i^b(j) > \tau_i^b(k)\}|, \tag{5}$$

which can be normalized by its largest value $n(n-1)/2$, where the n is the length of τ_i^m. Thus a larger Kendall distance indicates a larger diversity. The average and the distribution of the normalized Kendall distance are illustrated in Figure 6. They indicate that *more than half of the apps are not in the same order in two recommendations*, which further identifies the significant diversity within two methods.

4.3 Diversity on Network Level

App recommendations eventually form a recommending network in the online market, which influences the users directly on their online behaviors. Therefore, we further conduct measurements on the level of recommending networks.

Link Reciprocity. Firstly we investigate the link reciprocity, which captures a bi-node relationship in the network. It is defined as the proportion of its reciprocal links:

$$r = |\mathcal{E}^{\leftrightarrow}|/|\mathcal{E}|, \tag{6}$$

where $\mathcal{E}^{\leftrightarrow}$ represents the reciprocal links and $\mathcal{E}$ is the edge set. Since a reciprocal link denotes that two edges connect a pair of nodes from both directions, the link reciprocity of the networks indicates the tendency of nodes to form mutual connections between each other.

In the recommendation networks, a reciprocity link indicates that two apps are recommended to each other so that users then can jump forward and back across web pages between them. Reciprocal links thus facilitate the surfing convenience for users. Meanwhile, since there is limited space on web pages to recommend apps, the reciprocal links on the other side would occupy the chances for other apps to be presented. The results in 3 shows that there is *a slightly larger link reciprocity of MA recommendations.* This suggests that MA apps are connected more closely thus they may exhibit a better surfing convenience but lower app discovery efficiency.

Clustering Coefficient. We then examine the clustering coefficient of recommending networks on the triple node level. It measures the degree to which the neighbors of a node are connected. This tendency has been observed in most real-world networks, especially in social networks [5]. Particularly, in the recommending networks, the clustering coefficient captures whether apps recommended to the same app are also recommended to each other. A larger clustering coefficient suggests that the recommended apps are connected more closely thus maybe more similar to each other. However, the recommendations that are more cohesive also result in more efforts for users to discover novel apps.

We take the definition of clustering coefficient around a node i as the number of triangles in which the node i participates, normalized by the maximum possible number of such triangles:

$$c(i) = 2\mathcal{T}_i / d_i(d_i - 1). \tag{7}$$

The $\mathcal{T}_i$ is the number of triangles through the node i and d_i is the degree of the node i. The clustering measurements in Table 3 show *a larger clustering coefficient in the MA recommendations.* It indicates that apps recommended by MA method are connected more closely to each other, which may results in recommending more similar apps while restricting the app discovery scope.

Strongly Connected Component. On the multi-node level, we investigate the SCC (strongly connected component) in the recommending networks. SCC in a network is the part of network in which every node can be reached by another. In the context of app recommendations, each app in the same SCC of the network can be discovered by another. Therefore, a smaller number of the SCCs may result in a better connected network, so that users can reach more apps from one of them. In addition to examining the number of SCCs, we also illustrate the distribution of the SCC sizes in each recommending network in Figure 7. From the illustration we see that the *MA recommendations are more dispersed on the network level, while HB recommendations form less SCCs and a better connected network.*

Overall Observations on Network Level. Measurements on the network level collaboratively suggest that the MA method provides recommendations that are more cohesive thus the recommended apps are connected more closely. This would help users surfing more conveniently on the web pages while discovering apps that are more similar. On the other side, the HB recommendations have formed a recommending network which is less dispersed so that users may reach more apps from one of them in the network.

5 Conclusion

In this paper, we firstly propose a new recommending method based on the analysis of app metadata. We then measure the diversities within our method and the existing method from various levels. Through the measurements we obtain new knowledge of user preferences and gain better understanding of the CF and CB approaches in the context of recommending mobile apps. We further indicate the necessity and potential of evolving the existing mobile app recommender systems. Thus lots of future studies can be done motivated by our work, such as the app similarity measurement, the modeling of mobile users, and the design of optimal hybrid recommending schemes.

References

1. Cnet: Google ties apple with 700,000 android apps (April 2013), `http://news.cnet.com/8301-1035_3-57542502-94/google-ties-apple-with-700000-android-apps/`
2. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41, 391–407 (1990)
3. Gemmis, M.D., Iaquinta, L., Lops, P., Musto, C., Narducci, F., Semeraro, G.: Preference learning in recommender systems. In: ECML/PKDD 2009 (2009)
4. Newman, M.E.J.: Mixing patterns in networks. Physical Review E 67(2), 026126 (2003)
5. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (1998)

An Urban Area-Oriented Traffic Information Query Strategy in VANETs*

Xinjing Wang[1], Longjiang Guo[1,2,**], Chunyu Ai[3],
Jinbao Li[1,2], and Zhipeng Cai[4]

[1] School of Computer Science and Technology, Heilongjiang University, China
[2] Key Laboratory of Database and Parallel Computing, Heilongjiang, China
[3] Division of Math & Computer Science, University of South Carolina Upstate, USA
[4] Department of Computer Science, Georgia State University, Atlanta, USA, 30303
longjiangguo@gmail.com, aic@uscupstate.edu, zcai@cs.gsu.edu

Abstract. Traffic information query in Vehicular Ad Hoc Network has various significant applications. Real-time traffic information can provide support for users to choose an optimal route according to current traffic situation. In this paper, we propose an urban area-oriented traffic information query processing mechanism, which can acquire the real-time traffic information of multiple paths from source to destination in relatively fast and accurate manner, and help users to determine an optimal route. The proposed mechanism includes two key algorithms - query dissemination and processing, and routing results backward to query requester. The query processing algorithm determines the scope of each query, so that a vehicle can query and collect data within a certain efficient scope to avoid returning overwhelmed large amount results. For queried vehicles, returning results to the moving query requester is a dynamic routing problem. We proposed a position predicting method to estimate the current location of the requester according to the information stored in the query packet. Simulation results show that the proposed strategy can improve the efficiency of data transmission, and the returned query results is effective for choosing an optimal route.

* This work is supported by Program for New Century Excellent Talents in University under grant No.NCET-11-0955, Programs Foundation of Heilongjiang Educational Committee for New Century Excellent Talents in University under grant No.1252-NCET-011, Program for Group of Science and Technology Innovation of Heilongjiang Educational Committee under grant No.2013TD012, the Science and Technology Research of Heilongjiang Educational Committee under grant No.12511395, the Science and Technology Innovation Research Project of Harbin for Young Scholar under grant No.2008RFQXG107, 2009RFQX080 and No.2011RFXXG014, the National Natural Science Foundation of China under grant No.61070193, 60803015, Heilongjiang Province Founds for Distinguished Young Scientists under Grant No.JC201104, Heilongjiang Province Science and Technique Foundation under Grant No.GC09A109, The Natural Science Foundation of Heilongjiang Province under Grant No. F201038.

** Corresponding author.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 313–324, 2013.

1 Introduction

The communication framework built for Vehicular Ad hoc Networks(VANETs) is significant foundation of the Intelligent Transportation System(ITS). The wireless communicating technology of VANETs can be used to implement applications in transportation field such as traffic accident prediction, driving assistance, and traffic information query. Also, it can provide services such as data download, vehicular entertainment, and data query combined with WiFi. It can be foreseen that VANETs will provide broader space for developing the traffic intelligent and humanized improvement, thus making traveling more convenient.

A good driving route planner can help scheduling routes so as to save time and energy consumption and avoid traffic congestion. The traffic information query algorithm proposed by this paper is a necessary support of optimal path computation. To ensure timely updating data, the algorithm shrinks the query region and possibly divides a query to sub-queries. In order to refine data redundancy, our algorithm removes data of paths that obviously are not candidates of the best path. To return results,a datum returning routing algorithm is proposed to predict the current location of the query requester vehicle. Simulation results show that the strategy can help improve the efficiency of data transmission, and the returned data is relatively effective.

The rest of this paper is organized as follows. In Section 2, we introduce related works. The design and analysis of traffic information query algorithm are addressed in Section 3. Analytical performance study and simulation results are presented in Sections 4. Finally, we conclude this paper in Section 5.

2 Related Work

Data management in VANETs is a popular research issue at present, and so far there are a lot of research achievements in many aspects. Some parallel studies[24][25] also have arisen in wireless sensor networks. However, they are not suitable for VANETs. Y. Huang et al. [26] mainly analyzed multi-cast capacity for social-proximity urban bus-assisted VANETs. L. Wang et al. [6] studied the data naming issues during broadcast, and Hsu-Chun Hsiao et al. [7] focused on the issue of data signatures. N. Lu et al. [8] analyzed data throughput and delivery delay in VANETs in detail. In recent years, this field has risen to a hot research topic: V2G (Vehicle to Grid). Due to the popularity of today's electric vehicles in the city, the connection between VANETs and the smart grid becomes significant. The preliminary assumption is that the grid can sell electricity to PHEVs (plug-in hybrid electric vehicles) through the V2G network, and PHEVs also can sell their excess electricity back to the grid, resulting in energy saving and a win-win situation on economic benefits. Even though there are more and more studies in V2G, the majority of these studies still stay at a theoretical level, yet they make fairly strong foundation for future research. For instance, Albert Y. S. Lam et al. [9] established an economic model to specify the working principle of the V2G, thus making an intuitive foundation for V2G research; Romain Couillet et al. [10] also made a formal description of the mean field equilibrium of the resulting competitive interaction when EV (electric vehicle) owners buy and sell electricity from their cars, selfishly but rationally, based on collective price incentives with economic models; Y. Li [11] provided an energy efficient solution to solve a problem about the integration of PHEVs in smart grid with renewable energy. In this paper we study another data management issue, which is urban traffic information data query.

In VANETs, data transmission routing algorithms are broadly divided into five categories according to routing mechanisms [12], respectively are based on connection [13], based on mobility [14], based on infrastructure [15], based on geographic position [16], and based on probability model [17]. In addition, ETP [18] discussed the technical feasibility of using DSRC wireless communication devices to achieve opportunity communication during the process of two moving vehicles encounter; Y. Wu et al. [19] calculated and predicted travel trajectory of vehicles using Markov chain, and builds the routing protocol according to it. In this paper, we consider that the transmitting destination is always moving during data transmission, so we design an algorithm by predicting the location of a destination vehicle so as to route query results back to the query requester.

3 Traffic Query and Data Aggregation Strategy

The application background of our proposed method would be the city traffic scene. Traveling vehicles in a city are abstracted to a mobile model. If a vehicle v sends a query to request a route to a destination d, we use s to represent the current location (departure location) of v, the query issuing time is t, and the communication range between vehicles is R. In order to determine the best driving route, the vehicle v acquires traffic information of reachable paths from s to d by sending queries to other vehicles within its transmission range. Since the traffic may change at any time, query returned data can be effective only when the requester vehicle receives the desired data in a short period of time.

Due to the fact that a city can be big, the distance from starting point to the destination may be far, so that if the query requesting vehicle waits until receiving all data in the region, it will take long time. Long query respond delay can affect the traveling efficiency. Also, since traffic information is time sensitive, long respond delay might cause expired information being used thus leading to poor performance. Besides, not all paths from the source to the destination are eligible to participate in the evaluation of the best path, because the distances of some paths are too long compared with others to be considered regardless of road conditions. Therefore, these useless paths need to be excluded from queries in order to reduce the size of query packets and data redundancy of returned results. Moreover, when a vehicle receives a query, if it stores data needed by the query requester vehicle, it will transmit the data back to the requester; however, at that time the requester may have moved to another location, so routing the query results back to the requester is an issue need to be addressed. In Section 3.1 and 3.2, we will introduce how our proposed method solves these two problems.

3.1 Traffic Information Query Processing Algorithm

Related Definition and Proposition. Currently GPS can show all the reachable paths from starting point s to the destination d. Some navigation systems also can provide real-time traffic information and automatically suggest alternate routes when an accident has occurred along the route. However, these kind of traffic data services cannot guarantee the integrity and real time of traffic data. Our method generates a graph which includes all reachable paths from s to d by GPS. The desired data of a query are the traffic information of these paths.

The purpose of querying traffic information is to choose the best path to the destination which can balance both time and energy consumption. Even though the distance of a path is not the only determinant, it is still an important factor to limit the searching scope of a query. Because both time and energy consumption are considered, a very smooth path might not considered if it is a lot of longer than other paths.

A vehicle travels from location $(x_s,\ y_s)$ to $(x_z,\ y_z)$. If it collects the traffic information of all reachable paths from $(x_s,\ y_s)$ to $(x_z,\ y_z)$, it will receive so many redundant and useless data. These redundant and useless data are not necessary for choosing the best path. Therefore, a constraint can be added to a query to avoid generating so many redundant and useless data, that is, excluding regions that don't include the best path obviously.

If coordinates of the starting point s of a traveling vehicle are $(x_s,\ y_s)$, and coordinates of the destination are $(x_z,\ y_z)$ then any point (x, y) in the query searching region of path information satisfies

$$\begin{cases} (min(x_s, x_z) \leq x \leq max(x_s, x_z))\ or & \\ (min(y_s, y_z) \leq y \leq max(y_s, y_z)) & (x_s \neq x_z)\ and\ (x_s \neq x_z) \\ min(x_s, x_z) \leq x \leq max(x_s, x_z) & y_s = y_z \\ min(y_s, y_z) \leq y \leq max(y_s, y_z) & x_s = y_s \end{cases} \tag{1}$$

An intuitive judgment of a good path is that the path does not go away from the destination, such as the shaded area in Fig.1. If a vehicle travels in the non-shaded area, that is relatively getting further from the destination, the distance of a path like this should be longer than others within the shaded area. Even if the traffic information of these kind of paths are collected back, they will not be chosen to be the best path because of the longer distance. In order to improve efficiency and avoid transmitting useless data in the network, a query search scope is set according to the above formula.

If $(x_s \neq x_z)$ and $(y_s \neq y_z)$, the scenario is shown in Fig.1. The shaded area is union of the region between two x coordinates and the region between two y coordinates of the source and destination. If $y_s = y_z$, the region is bounded by x coordinates of the source and destination as shown in Fig.2. When $x_s = x_z$, the region is bounded by y coordinates of the source and destination.

When a vehicle receives a query, it will forward the packet to its neighbors. Since in urban areas, the starting point s may be very far away from the destination z, in order to avoid flooding the query to the whole network and guarantee

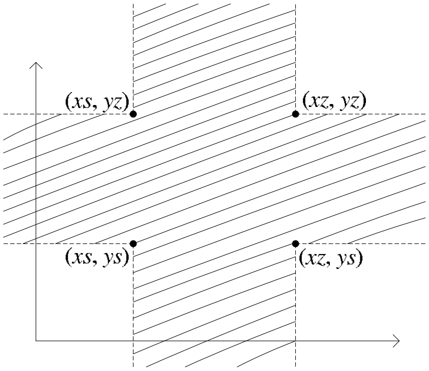


Fig. 1. $(x_s \neq x_z)$ and $(y_s \neq y_z)$

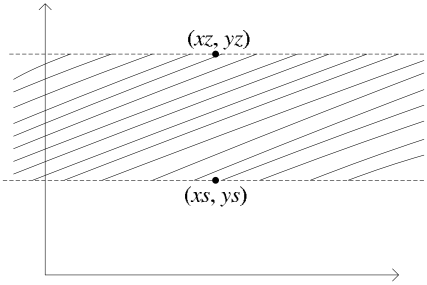


Fig. 2. $y_s = y_z$

the returned results fresh, we need to calculate in blocks, which means collecting data in k hops in a short time to ensure its effectiveness. Therefore, a query packet is transmitted at most in k hops. Vehicles along the potential best path might not receive the query and report the traffic information. To solve this, we pick a *path transit point* as the current destination. The location of the next path transit point is determined according to returned results from k hop neighbors. A location closer to the destination and with a fast path from the requesting vehicle is preferred. The distance between the source and the next path transit point must be k hop reachable. When the vehicle almost arrives the destination by following the best path generated by the algorithm, it sends out the query again to obtain the best path to the next path transit point. The process is repeated until the query can reach the destination area directly. An example is shown in Fig.3 where $Z1$ and $Z2$ are path transit points.

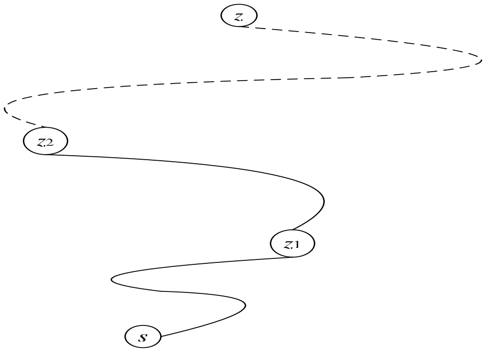


Fig. 3. Path transit point

Lemma 1. *The distance between the starting point of a query and the furthest location the query can reach is kR.*

Proof. The communication range of vehicles is R, so the longest distance a packet can travel in one hop is R. As mentioned above, if the destination is to far to be reached, the traffic information within k hops are collected to be used to determine the next path transit point. Therefore, the furthest distance a query can travel is kR; in other words, the longest distance between the query start point and a query receiver is kR. ■

A query packet includes: query requesting vehicle ID, starting point $s(x_s,\ y_s)$ and destination $z(x_z,\ y_z)$, hop count k, the current velocity v of the vehicle, and the time that the vehicle issues the query t.

Algorithm Description. A GPS equipped on vehicles records all paths in the city, and these paths form a map. We abstract the map to a graph by converting an intersection to a vertex and a road to an edge. The query transmitting mechanism is describe as follows.

Query requesting vehicle records reachable paths from the starting point s (x_s, y_s) to the destination z (x_z, y_z) within the searching scope defined in Formula 1. These paths are sorted by distance and encapsulated into the query packet. The query packet is transmitted in z's direction. If returned results are insufficient to determine the best path to z, a path transit point is chosen and the best path to the path transit point is generated according to the returned traffic information. This is repeated until reach the destination.

When a vehicle receives a query packet, it checks whether its current location is in the query list, if so, it returns the traffic information to the query requesting vehicle by using the Backward-Rule algorithm described in the next section. Then, it deletes that road from the query packet, sets the hop count to $k-1$ and forward the packet in z's direction. If the hop count is 0, it would not forward the query packet.

3.2 Returning Results

When a vehicle receives a query packet, it checks whether the query requests the traffic information of its current location, if so, it generates a packet containing the traffic information and sends it back to the query requesting vehicle. Since the query requester keeps moving after it sent the query, it is difficult to locate it in accurate and timely manner. Therefore, how to route the result packet back to it is an issue. Our strategy forecasts the current location of the query requesting vehicle via travel trajectory prediction.

V_t is the vehicle which needs to transmit the query result; V_r is the vehicle which issues a query and needs to receive query results. The knowledge V_t can obtain from the query packet is the location coordinates of the query requesting vehicle at the time t where t is the time when the query was sent out.

Lemma 2. *If the hop count stored in the query packet is* k' *when* V_t *receives the packet, the result packet needs to be transmitted at most* $k_b = k - k' + 1$ *hops to reach* V_r*.*

Proof. If the hop count is *k'* when the query packet is received by V_t, the packet has been transmitted (k - k'+1) hops. So the longest distance from the receiver to the sender V_r's original location is (k - k'+1)R according to Lemma2. Therefore, in $k - k' + 1$ hops, the result dat packet can be transmitted back to the original location of V_r.

While V_r keeps moving towards the destination after it sent the query and during query dissemination, processing, and transmitting returning results. Since V_t is on a path from V_r to the final destination, V_r also moves forwards V_t. So that the distance V_t and V_r's current location must be less than (k - k'+1)R, thus the data packet can arrive V_r in (k - k'+1) hops. ■

The reason of limiting the hop count to a small number is to reduce the transmission redundancy in the network. Less transmission workload can decrease the probability of channel conflicts and collisions.

Algorithm Description. When V_t wants to return the data back to V_r, it will predict the current location of V_r roughly according to the information in the received query packet, and sending data back to V_r by using the Backward-Rule Algorithm as shown in Algorithm 1. V_t calls Backward-Rule algorithm to decide transmit the result packet to which neighbors. If a receiver of the query result packet is V_r, V_r stores the result locally and waits to receive all necessary traffic

Algorithm 1. Backward-Rule Algorithm

```
Input: Time in query packet $t$; Current time $t_n$; ID of $V_r$; ID of $V_t$; Coordinates of $V_r$ $(x_s, y_s)$; Coordinates of $V_t$ $(x_t, y_t)$; $v$ is the maximum speed of vehicles.
if $k_b \neq 0$ and ID of $V_r \neq ID\ of\ V_t$ then
    $k_b = k_b - 1$.
    Abstract all the reachable paths from $(x_s, y_s)$ to $(x_t, y_t)$ to a graph $G$.
    Delete the paths which are not in the searching scope defined in Formula 1.
    $S = v \times (t_n - t)$.
    for each edge $E_{ij}$ in Graph $G$ do
        if $V_r$ may arrive $E_{ij}$ (judging according to $S$ and $V_r$'s moving direction) then
            Insert $E_{ij}$ to $Road$.
    for each road segment $Road[m]$ in $Road$ do
        Generate a vector between $(x_t, y_t)$ and $Road[m]$, store it in $Vector[m]$.
    for each $Neighbor[m]$ in $Neighbor$ do
        for each $Vecotr[n]$ in $Vector$ do
            if The vector from $(x_t, y_t)$ to $Neighbor[m]$ and $Vector[n]$ are in the same direction then
                $V_t$ transmits the data packet to $Neighbor[m]$.
```

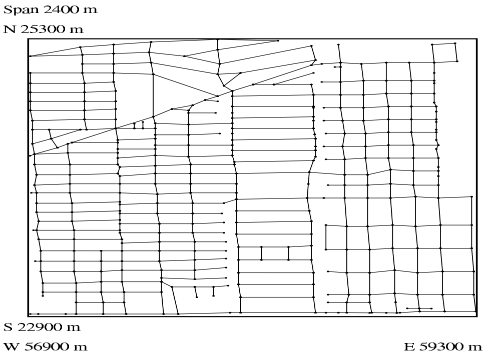


Fig. 4. Simulation Scenario

information to choose an optimal route. If the receiver is not V_r and k_b is not 0, it runs Backward-Rule algorithm to determine transmit the result packet to which neighbors. Neighbors of V_t which are in the direction of V_r's possible current locations are preferred to be chosen to forward the result packet.

In Algorithm 1, S is the estimated distance that V_r has traveled since it sent the query out; *Road* is an array storing the edges (as mentioned in Section 3.1, the map is converted to a graph); *Vector* is an array of vectors from (x_t, y_t) to a road segment *Road*$[m]$; *Neighbor* is an array storing the current neighbors of V_t; E_{ij} is an edge which connects vertex i and j, that is, a road segment which connects intersection i and j.

4 Simulation Experiments

4.1 Experimental Environment Settings

In the experimental environment setting, vehicles are traveling within a $3km \times 3km$ fixed region of the street environment. We implement our method by NS2 [20] to evaluate the performance. In order to provide a real VANETs environment, we use SUMO simulator[21] and a street map in Tiger database[22] (as shown in Fig.4) to generate a street topology model and vehicle moving trajectory file, and the file is the input of our simulator.

In the moving scenario, there are 2,000 vehicles, Vehicles's maximum velocity is 70km/h, and R is 150m.

4.2 The Simulation Comparison Schemes

In the proposed strategy, there are two key algorithms: data query algorithm and returning results algorithm. In the data query algorithm, even though the query packet is transmitted in a certain range, it is still a broadcast algorithm essentially. Therefore, we compare our proposed algorithm (named *Strategy*) with Epidemic [23], Scheme 1, and Scheme 2, where Scheme 1 is same as our

Strategy except without using searching scope, and Scheme 2 is also same as our Strategy except using flooding to return results instead of Backward-Rule algorithm. Table1 shows the technologies used by each scheme.

Table 1. Technologies used in Schemes

Scheme	Transmit directive	Limit query range	Aggregate with Backward-Rule
Epidemic		√	√
Scheme 1	√		√
Scheme 2	√	√	
Strategy in paper	√	√	√

4.3 Strategy Performance Comparison

Data Collection Efficiency. The data collection efficiency refers to the number of received result data, which is affected by many factors. First of all, when a vehicle sends a traffic information query, if there is no certain directional degree of transmitting the query, there will be a lot of query packets having no answer. However, these useless packets will make collisions with other packets in the network, and cause failure of some important packet transmissions, thus decreasing the efficiency of data collection. In addition, when a piece of useful traffic information is found, if the speed of returning results is slow, the data is likely outdated when reaching the query requesting vehicle. Fig.5 displays the data collection efficiency of the four schemes. The data collection efficiency is the number of data obtained in unit time.

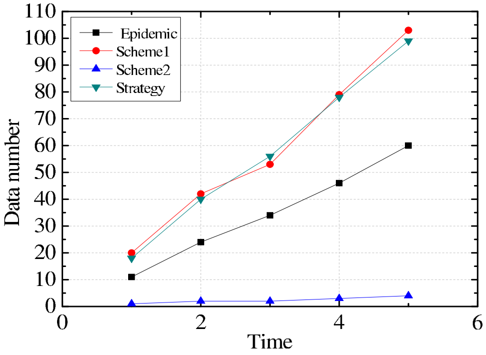


Fig. 5. Data collection efficiency

As seen in Fig.5, the efficiency of data collection of Epidemic is very slow because the query packets sent by Epidemic have no clear direction. However, the efficiency of Scheme 2 is also lower since Scheme 2 does not have a good result returning principle.

Data Collision Rate. Another important criteria is the data collision rate. The data collision rate is defined as the percentage of losing packets versus. Data collisions are the consequences of network congestion, and will result in packet loss. Fig.6 compares the collision rate of the four schemes. As shown in the figure, the collision rate of Epidemic and Scheme 2 is very high. From a global point of view, the sending packets method in Epidemic may cause the spread of many useless packets in the whole network, and these packets are redundant which cause collision with other packets. It may cause loss of packets and resulting in high collision rate. The problem in Scheme 2 is that it does not have a clear purpose in returning result, which also generates many redundant useless packets.

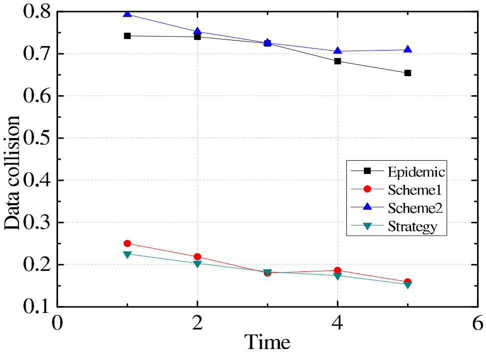


Fig. 6. Data collision rate

Proportion of Valid Data. The proportion of valid data is the number of the valid data received divided by the number of all received data. Data validation is related to the efficiency of data collection, especially the speed of returning results. We consider the returned data is valid if it is generated in the last t_v time duration where t_v is a user specified threshold and the coordinates of the sender is in the searching scope. Fig.7 shows the comparison between the proposed strategy and other methods.

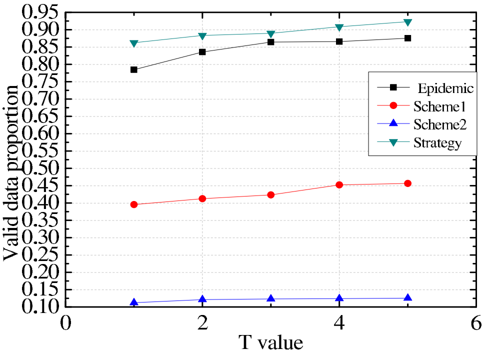


Fig. 7. Proportion of valid data

As shown in the figure, the valid data rate of Scheme 1 also returns useless data not in the searching scope. The reason that the valid data rate of Scheme 2 is lower because data don't have a clear direction to route the packet back to the query requester.

In summary, our proposed strategy has better performance compare with existing methods.

5 Conclusion

We proposed a real time traffic information query processing strategy to provide optimal route navigation in urban area. The proposed algorithm limits the searching scope to not only improve the efficiency but also help reducing network redundancy. A result returning algorithm, Backward-Rule, also is well designed to reduce network communication workload. Simulation results show that the strategy can achieve the desired effect.

References

1. Jeong, J., Guo, S., Gu, Y., He, T., Du David, H.C.: TSF: Trajectory-Based Statistical Forwarding for Infrastructure -to-Vehicle Data Delivery in Vehicular Networks. In: Proc. of IEEE, ICDCS 2010, pp. 557–566 (2010)
2. Zheng, Z., Lu, Z., Sinha, P., Kumar, S.: Maximizing the Contact Opportunity for Vehicular Internet Access. In: Proc. of IEEE, INFOCOM 2010, pp. 1109–1117 (2010)
3. Wu, Y., Zhu, Y., Li, B.: Infrastructure-assisted routing in vehicular networks. In: Proc. of IEEE, INFOCOM 2012, pp. 1485–1493 (2012)
4. Liu, N., Liu, M., Lou, W., Chen, G., Cao, J.: PVA in VANETs: Stopped cars are not silent. In: Proc. of IEEE, INFOCOM 2011, pp. 431–435 (2011)
5. Ahn, J., Wang, Y., Yu, B., Bai, F., Krishnamachari, B.: RISA: Distributed Road Information Sharing Architecture. In: Proc. of IEEE, INFOCOM 2012, pp. 1494–1502 (2012)
6. Wang, L., Wakikawa, R., Kuntz, R., Vuyyuru, R., Zhang, L.: Data naming in Vehicle-to-Vehicle communications. In: Proc. of IEEE, INFOCOM 2012, pp. 328–333 (2012)
7. Hsiao, H.-C., Bai, F.: Flooding-Resilient Broadcast Authentication for VANETs. In: Proc. of ACM, MobiCom (2011)
8. Lu, N., Luan, T.H., Wang, M., Shen, X., Bai, F.: Capacity and delay analysis for social-proximity urban vehicular networks. In: Proc. of IEEE, INFOCOM 2012, pp. 1476–1484 (2012)
9. Lam, A.Y.S., Huang, L., Silva, A., Saad, W.: A multi-layer market for vehicle-to-grid energy trading in the smart grid. In: Proc. of IEEE, INFOCOM 2012, pp. 85–90 (2012)
10. Couillet, R., Perlaza, S.M., Tembine, H., Debbah, M.: A mean field game analysis of electric vehicles in the smart grid. In: Proc. of IEEE, INFOCOM 2012, pp. 79–84 (2012)
11. Li, Y., Kaewpuang, R., Wang, P., Niyato, D., Han, Z.: An energy efficient solution: Integrating Plug-In Hybrid Electric Vehicle in smart grid with renewable energy. In: Proc. of IEEE, INFOCOM 2012, pp. 73–78 (2012)

12. Yan, G., Mitton, N., Li, X.: Reliable Routing in Vehicular Ad Hoc Networks. In: ICDCS Workshops 2010, pp. 263–269 (2010)
13. Abedi, O., Fathy, M., Taghiloo, J.: Enhancing aodv routing protocol using mobility parameters in vanet. In: Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications (AICCSA 2008), pp. 229–235. IEEE Computer Society, Washington, DC (2008)
14. Taleb, T., Sakhaee, E., Jamalipour, A., Hashimoto, K., Kato, N., Nemoto, Y.: A stable routing protocol to support its services in vanet networks. IEEE Transactions on Vehicular Technology 56(6), 3337–3347 (2007)
15. Kim, H., Paik, J., Lee, B., Lee, D.: Sarc: A street-based anonymous vehicular ad hoc routing protocol for city environment. In: Proceedings of the IEEE/IFIP International Conference on Embedded and Ubiquitous Computing (EUC 2008), pp. 324–329. IEEE Computer Society, Washington, DC (2008)
16. Kato, T., Kadowaki, K., Koita, T., Sato, K.: Routing and address assignment using lane/position information in a vehicular ad hoc network. In: Proceedings of the IEEE Asia-Pacific Services Computing Conference (APSCC 2008), pp. 1600–1605. IEEE Computer Society, Washington, DC (2008)
17. Yan, G., Olariu, S., Salleh, S.: A probabilistic routing protocol in vanet. In: Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia (MoMM 2009), Kuala Lumpur, Malaysia, December 14-16 (2009)
18. Yu, B., Bai, F.: ETP: Encounter Transfer Protocol for opportunistic vehicle communication. In: INFOCOM 2011, pp. 2201–2209 (2011)
19. Wu, Y., Zhu, Y., Li, B.: Trajectory improves data delivery in vehicular networks. In: INFOCOM 2011, pp. 2183–(2191)
20. The Network Simulator, http://www.isi.edu/nsnam/ns
21. http://sourceforge.net/apps/mediawiki/sumo/index.php?title=Main_Page (EB/OL) (2010)
22. Bureau, U.C.: Tiger, tiger/line and tiger-related products. Vahdata
23. Vahdata, A., Becker, V.D.: "Epidemic routing for partially connected ad hoc networks", Technique Report, CS-2000-06
24. Cai, Z., Ji, S., Li, J.: Data cachingCbased query processing in multiCsink wireless sensor networks. International Journal of Sensor Networks 11(2), 109–125 (2012)
25. Ai, C., Guo, L., Cai, Z., Li, Y.: Processing area queries in wireless sensor networks. In: Mobile Ad-hoc and Sensor Networks, MSN 2009 (2009)
26. Huang, Y., Guan, X., Cai, Z., Ohtsuki, T.: Multicast Capacity Analysis for Social-Proximity Urban Bus-Assisted VANETs. In: The IEEE International Conference on Communications, ICC 2013 (2013)

Navigation for Indoor Mobile Robot Based on Wireless Sensor Network

Yunzhou Zhang, Shuo Wang, Guanting Fan, and Jixian Zhou

College of Information Science and Engineering, Northeastern University,
P.O. Box 128, Shenyang 110819, P.R. China
zhangyunzhou@ise.neu.edu.cn, WangShuo8899@Gmail.com,
{532827215,471363855}@qq.com

Abstract. A practical system is proposed to solve the navigation problem for indoor mobile robot based on wireless sensor network (WSN). The discrete data acquired by WSN is processed to form a three-dimensional global topographic map, which is then converted into a 0-1 grid map through binarization. The grids where obstacles locate are expanded according to specific criteria to construct the robot route network. Then, the route-network-grid map is converted to directional weighted graph, with which the D*Lite algorithm can be used to solve the problem of shortest path between two fixed nodes and acquire the optimal node set to construct the optimal path. Simulation result shows that the indoor environment can be well expressed by the proposed modeling method, from which we can accomplish the navigation to lead the mobile robot arrive to destination with the shortest distance in a dynamic environment.

Keywords: Mobile robot, Navigation, Wireless sensor network, Indoor environment modeling, Grid map.

1 Introduction

Robot navigation is a fundamental problem in mobile robot research, it plays an important role in a wide variety of applications, especially for human existence detection in the case of disasters. It's very dangerous for first responders to search victims or investigate in the inside of buildings damaged by a natural or man-made disaster because such buildings may collapse further or contain harmful substances [1]. Using robots with navigation technology can replace manpower research and provide much safer and more accurate rescue. Robot navigation can be summarized the problem into answering the following three questions [2]: "Where am I?" "Where am I going?" and "How should I get there". The first question lies to identifying the current location of the robot. The second and third questions are related to the capability of environment perceiving and path planning. So how to resolve these questions and improve environment sensing capability have been a hot area of research.

In recent years, robot navigation based on wireless sensor network (WSN) is widely noted and investigated [3]. As a novel network technique, WSN has

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 325–336, 2013.

extensive prospects in battle field, environment supervision, household automation and many other fields [4]. Some characteristics of WSNs which are low cost, low power requirements, multifunction capabilities, robustness and scalability [5] greatly extend the sensing capability of the mobile robot and make it possible for the robot to move beyond its current-sensing range and respond to distant events.

In this paper, Robot collects data and information from the wireless sensor network and then constructs the real-time moving path according to the dynamic environment. Wireless sensor network provides mobile robots with the real-time global perception to monitor continuous and long-rage environment. This involves representation of world model and path search strategy, but with restricted adaptability, the robot cannot deal with environment change or dynamic obstacle. As a result they bring forward high demand for the quality of planning method and algorithms. Through effectively integrating global planning and environment modeling, this paper proposes a novel navigation system based on wireless sensor network using grid map and expanded method. It provides a new choice for motion planning of mobile robots in complex indoor environment.

The rest of the paper is organized as follows. Section 2 provides a brief overview of the navigation of the mobile robot based on WSN. In section 3 we introduce some basis of sensing model of ultrasonic. We will give the details of mechanism of how the WSN equipped with ultrasonic to construct the dynamic indoor environment in section 4. The D*Lite algorithm and the simulation and its experiment results will presented in section 5. The conclusions are given in Section 6.

2 Related Works

For mobile robots in WSN field, path planning related research can be divided into three directions such as grid map, intensity and potential field and information broadcasting. The most frequently used method is grid map, where an occupancy grid map of the immediate surroundings of the robot is created. It is used to determine the navigation direction such that the robot is safely guided towards a goal. In [6], the author introduces a solution for the complex problem of autonomous map building with an incremental grid based mapping technique that is suitable for real-time obstacle detection and avoidance. In [7], WSN is used to acquire the real-time information of surrounding environment. Then environment information model and grid map are built to provide decision support for mobile robots. To decrease the communication traffic, reference [8] used grid map to describe the physical connection between environment situation and the robot. It can obtain minimal communication bandwidth and energy cost when approximate optimal path is acquired.

The potential field can be built according to the signal intensity and form navigation area for mobile robot. In [9], the sensor network models the danger levels sensed across its area and has the ability to adapt to changes. It represents the dangerous areas as obstacles. In [10], the coordinate system, which denotes the mobile robot's state and navigation space, is set up based on the RSSI potential field. Each beacon node can detect the distance which is quantized by RSSI value and calculate the control outputs. The author [11] addresses a novel method called DVFF combining the Virtual Force Field (VFF) obstacle avoidance approach and global path planning based on D* algorithm. Some researchers think that the navigation path can be built for mobile robots through the information spreading in wireless sensor network. In [12], an algorithm is presented that allows the robot to navigate precisely and reliably using a deployed sensor network embedded in the environment. Sensor nodes act as signposts for the robot to follow, thus obviating the need for a map or localization on the part of the robot. Similarly, a sharing network storage mechanism is used in [13] to build environment sensing model. Robot can obtain the global coordinate via ultrasonic signal when moving. For robot path planning, the aim is always to obtain the shortest and safest path at small cost of communication and calculation. One way to calculate a robot's route is to adopt existing routing algorithms, such as DSR and AODV, used in the ad hoc network [14]. Reference [15] proposes a distance-aware robot routing (DAR) algorithm, where the shortest path is selected by allowing a mobile robot to communicate with surrounding nodes through wireless communication.

Considering such factors as the data accuracy and real-time requirement, we choose the grid map to represent the surrounding environment. Then, barrier expansion is proposed to construct the mobile path network for mobile robot based on the surrounding environment information. Finally, we use the D*Lite algorithm in graph theory to plan the optimal path for mobile robot.

3 Sensing Model

Ultrasonic wave transmitted by sonar sensors is not a straight line, but owns an open-angle with certain directivity. Its energy is mainly concentrated around the central axis with the angle rage θ which is called the main lobe. While the energy in side lobe is relatively small. The energy distribution of ultrasonic sensor is shown in Fig. 1.

Each data returned from ultrasonic sensor is a real number, which is assumed to be S. The median line model provides the simplest explanation of this real number, where the obstacle located in the direction of angle 0° (median line) corresponds to the distance S. When the sensor is close to the obstacle and the angle of ultrasonic beam is small, the median line model is an easy and fast method to implement obstacle avoidance [8].

In this article, Cricket node is used to measure the environment dynamically. Under indoor environment, Cricket can measure the distance at about 5.5m with no direction deviation and about 3.8m with 30° direction deviation. When the distance is about 1∼3.5m and 3.5∼10m, Cricket's accuracy can attain about 1cm and 2cm respectively. Through reasonable configuration, there will be no disturbance among multiple Cricket nodes.

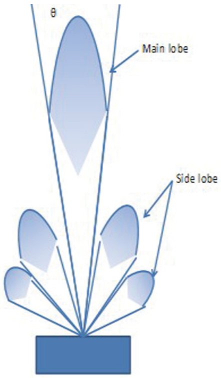


Fig. 1. Energy distribution of the ultrasonic sensor

4 Perception and Expression of Global Environment

4.1 Deployment of Wireless Sensor Network

The indoor environment is divided into $m \times n$ grids with the length of each grid's side as L. The sensor nodes are evenly deployed on the ceiling according to those grids. The center of each grid corresponds to a sensor node. Then we can build a wireless sensor network shown in Fig. 2. In order to create a map that can be updated dynamically in the variable environment.

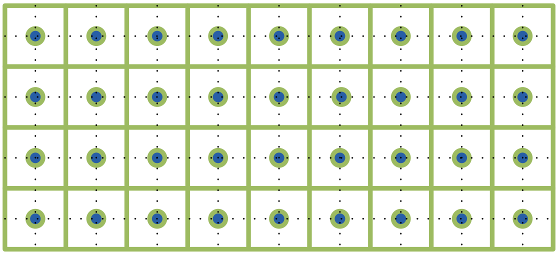


Fig. 2. Wireless sensor network deployment

Wireless sensor network nodes are used to monitor their surrounding environment periodically. Then, the system judges whether the difference between the returned value and that of last time exceeds the permissible error range. If the difference exceeds the range, WSN will re-establish the environmental map and provide real-time information for path planning.

4.2 Three-Dimensional Map of the Environment

The indoor environment is a 6m × 4.8m rectangle, while the maximum diameter of the mobile robot is about 24cm. We use median line model as our ultrasonic

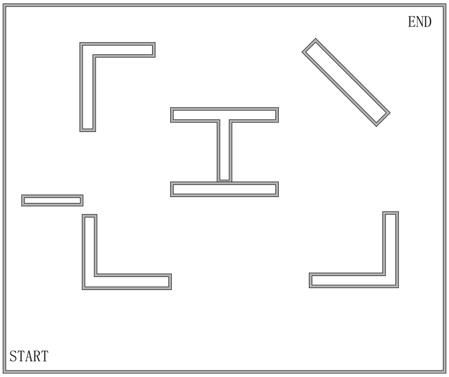


Fig. 3. Indoor environmental map

sensors model and make environmental modeling based on WSN information. Fig. 3 shows an actual layout of the indoor environment.

Take the actual diameter of mobile robot and a safe distance away from the obstacles into account, the map is divided into 20 × 11 square grids with side length as 12.8cm. The space rectangular coordinate system is established with origin of lower left corner vertex. Each grid will accommodate only one sensor node. Then, the wireless sensor network deployed on the room ceiling can be used to measure the distance from the sensor node to the obstacles. In fact, what we need is the exact height of the obstacles. Therefore, we could use the following formula to figure out obstacle's height according to the returned value.

$$h_{ij} = H - s_{ij} \quad (i = 1, 2, \cdots 11 \; j = 1, 2, \cdots 20) \tag{1}$$

Where H is the distance between the sensor node and the ground, $i \times j$ represents those 11×20 grids, s_{ij} is the value returned from sensor node corresponding to the grid, h_{ij} represents the barrier's height corresponding to the grid.

As Fig. 4 shows, the measured data of obstacles' height are discrete and clearly cannot meet the requirement of grid map construction.

In order to obtain the environment information of each location in the whole space, we need to establish the environment model. Therefore, we estimate the environment information of those blank locations, where the wireless sensor nodes cannot sense, by using the cubic spline interpolation method. Then we build the environment model of robot's working space. Cubic spline interpolation is a simple interpolation method which has good continuity. It owns high accuracy, none distortion and can achieve better convergence with low cost. Fig. 4 shows the measured discrete barrier height data. In order to guarantee small error, the interpolation method we adopt must have the characteristics of rapid decay. Otherwise, the range covered by the data in Fig. 4 will distorted seriously. Cubic spline interpolation exactly conforms to this condition. According to several experimental tests, cubic spline interpolation can guarantee the data

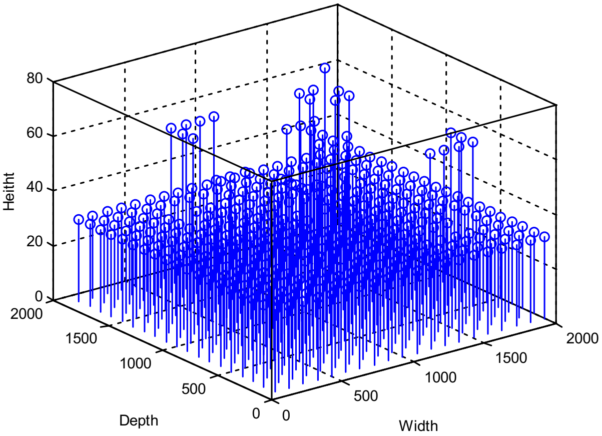


Fig. 4. Measured height data of obstacles

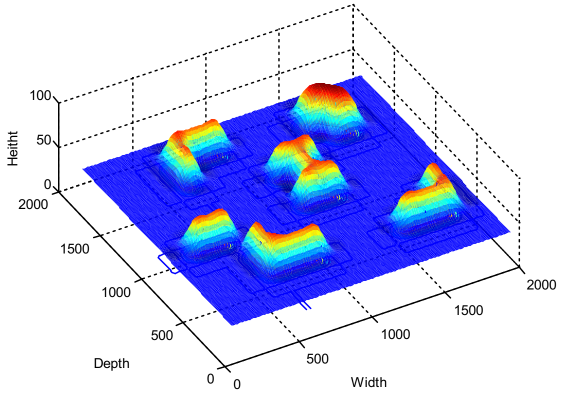


Fig. 5. Three-dimensional map of the global environment

beyond obstacle bound to decay quickly. This matches the actual situation for ultrasonic sensors to detect obstacles. The interpolated obstacles data is shown in Fig. 5.

Comparing Fig. 3 with Fig. 5, we can see that the interpolated 3-D map of global environment is close to the actual layout. The 3-D map can provide a better environment foundation for following path planning.

4.3 Establishment of Grid Map

In order to create a grid map which is easy to implement mobile robot navigation, the environment information is processed according to such factors as the 3D model, robot performance and structure. According to the result, the grids are assigned values to acquire binarized grid map. Considering the object features and robot performance, proper threshold is set up to judge whether the obstacle will hinder the robot movement. According to the performance of mobile robot and specific environmental characteristics, an appropriate threshold is set to judge whether the obstacles impede robot to march forward. Hence, when obstacle's height is above the threshold, the robot will be unable to pass through. Otherwise, it is believed that the mobile robot can move across the obstacle safely. By repeated tests, we find that the robot we adopt cannot pass through obstacles higher than 10cm.

The threshold of grid model can be expressed as follows:

$$f(x,y) = \begin{cases} 1 \text{ if } (h(x,y) \geq h_0) \\ 0 \text{ else} \end{cases} \tag{2}$$

Here, (x, y) is the coordinates of grid cell, $h(x, y)$ is the height value at grid (x, y) which is calculated by the sensor node, h_0 is a predefined threshold, $f(x, y)$ is Boolean value which represents the value on the environmental map at grid (x, y). Value 1 represents that obstacle exists, while 0 means no obstacle. Here, we suppose that when the height is over 10cm, that is, $h_0 = 10$, impendent effect will be formed for the robot's movement. That is, the robot is unable to pass through. Then we could get the grid map of height, as shown in Fig. 6.

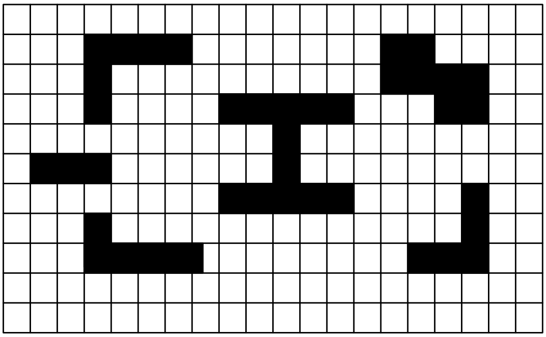


Fig. 6. Binarized grid map

Fig. 6 shows the grid map obtained after binarization based on the height of three-dimensional topographic map. ■ means the obstacle area, while □ means the area where the mobile robot can safely pass through.

5 Global Path Planning for Mobile Robot

5.1 Path Network Construction

Based on the global environmental modeling, the concept of obstacle "expansion" is proposed in this paper to construct the route/path network. We regard the grids occupied by obstacles and their boundaries as a "core", which will be gradually expanded to its adjacent free grid area. Each expansion happened in the working space is equal to a layer of growth for the obstacles. It will continue expanding until it meets the stop conditions.

- Obstacle expansion rules are defined as following:
  - At each time, the grids occupied by obstacles will grow uniformly in four directions including up, down, left and right.
  - The grids grown after each obstacle expansion are treated as new obstacles.
  - If the expansion satisfies the stop conditions in one direction, then in the next expansion, the grids stop expanding in this direction and continue expanding in other directions.
- Accordingly, the expansion stop conditions are defined as following:
  - Considering that the exact diameter of the robot is (2/3) L, the expansion will stop when the distance between two obstacles or between obstacle and the boundary equals to a grid.
  - If the distance between the boundary and two obstacles or the distance among three obstacles is one free grid (as shown in Fig. 7), the expansion will stop.
  - It will stop expanding if free grid doesn't exist in the whole working space.
  - When the obstacles expansion meets any one of the first two conditions, it will stop expanding in this direction. When the third condition is satisfied, all obstacles will stop expansion.

When obstacles expand, the intersection of two areas will form an intersecting line, while the intersection of three or more areas will form an intersection points. We define the intersecting line as the feasible path that robot can pass through. All intersecting lines will constitute a network consisting of candidate paths. The intersection points are the basic nodes of the path network.

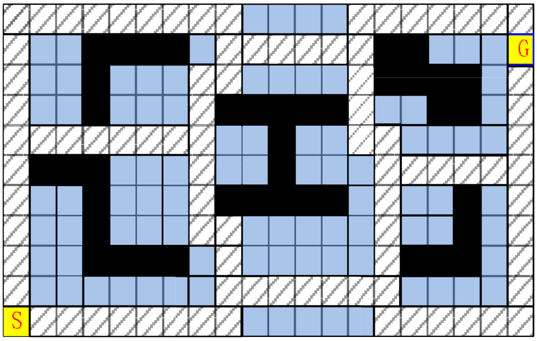


Fig. 7. Obstacle expansion area and path Network

As shown intuitively in Fig. 7, 🟦 is the obstacle expansion area, while ▨ is the path network for robot passing through safely. The grids marked number 1 to 10 are the basic network nodes.

5.2 Path Planning

In this way, the path network in Fig. 7 can be regarded as a weighted graph. To find out the shortest distance between fixed nodes in a weighted graph, we adopt D*Lite [16] as the path planning for the mobile robot in the context of the WSN.

*D*Lite path planning algorithm*

```
procedure CalcKey(s)
  return [min (g(s), rhs(s)) + h(sstart, s) + km; min(g(s), rhs(s))];
procedure Initialize()
  U $= \Phi$;
  km = 0;
  for all s $\in$ S rhs(s) = g (s) $= \infty$;
  rhs(Sgoal) = 0;
  U.Insert(Sgoal, CalcKey(Sgoal));
procedure UpdateVertex(u)
  if (u $\neq$ Sgoal) rhs(u) = mins$'$ $\in$ Succ(u) (c(u, s$'$) + g(s$'$));
  if (u $\in$ U) U.Remove(u);
  if (g(u) $\neq$ rhs(s)) U.Insert (u, CalcKey(u));
procedure ComputeShortestPath()
  while (U.TopKey()<CalcKey(Sstart) OR rhs(sstart) $\neq$ g(sstart))
    Kold = U.TopKey();
    u = U.Pop();
    if (Kold < CalcKey(u));
      U.Insert(u, CalcKey(u));
    else if (g(u) > rhs(u))
      g(u) = rhs(u);
      for all s $\in$ Pre(u) UpdateVertex(s);
    else
    g(u) $= \infty$;
    for all s $\in$ Pre(u) $\cup$ {u} UpdateVertex(s);
procedure Main()
  Slast = Sstart;
  Initialize ();
  ComputeShortestPath();
  while (Sstart $\neq$ Sgoal)
    /* if (g(Sstart) $= \infty$) then there is no known path */
    Sstart = arg mins$'$ $\in$ Succ(Sstart) (c(Sstart, s$'$)+ g(s$'$));
    Move to Sstart;
    Scan graph for changed edge costs;
    if any edge costs changed
      km = km + h(Slast, Sstart);
      Slast = Sstart;
```

```
for all directed edges (u, v) with changed edge cost
    Update the edge cost c(u, v);
    UpdateVertex(u);
ComputeShortestPath();
```

Fig. 8. D*Lite path planning algorithm

S denotes the finite set of vertices of the graph. Succ(s)∈Sdenotes the set of successors of vertex s∈S in the graph. Similarly, Pred(s)∈S denotes the set of predecessors of vertex s∈S in the graph. $0 < c(s, s') \leq \infty$ denotes the cost of moving from vertex s to vertex s'∈Succ(s). D*Lite always determines a shortest path from a given start vertex s$_{start}$∈S to a given goal vertex s$_{goal}$∈S, knowing both the topology of the graph and the current edge costs.

Fig. 9 is the effect when the optimal path projected in grid map. This sign "— ·▶" denotes the optimal path from start to goal calculated by D*Lite algorithm. If there is any change in the environment, the path will be adjusted promptly according to the new map.

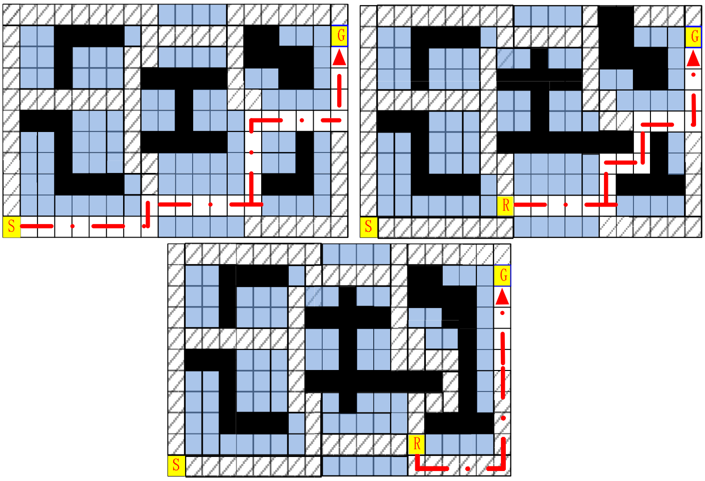


Fig. 9. Optimal path of grid map

5.3 Experiment Results

The experiment was conducted to prove the real-time performance. The computer equipment is Core i5 CPU 2.5 GHz. Some obstacles which are all squares are randomly put in the map. Figure 10 shows the analysis time of the algorithm compare with D*Lite without the assistant of WSN. The analysis is tested on hundred times to get the average value. The analysis rime ratio can prove our system is more efficient.

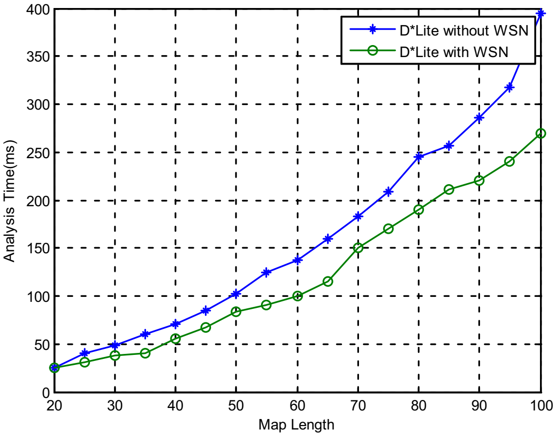


Fig. 10. The analysis time ratio

6 Conclusion

Through the ultrasonic sensing model, we analyzed and processed the information data acquired by WSN to build the grid map of indoor environment. Then, the path network is constructed via expansion method. The optimal path is established by using the method of D*Lite algorithm in graph theory. According to the simulation result, the proposed method can effectively give solutions to the problem of environmental modeling and path planning for mobile robots in indoor environment.

Acknowledgments. This paper is funded by National Natural Science Foundation of China (No. 61273078), the Open Foundation of State Key Laboratory of Robotics (No. RLO200913), China Postdoctoral Science Foundation (No. 2012M511164), Chinese Universities Scientific Foundation (No. N110404030, N110404004), Liaoning Doctoral Startup Foundation (20124004).

References

1. Kumar, V., Rus, D., Singh, S.: Robot and Sensor Networks for First Responders. Pervasive Computing 1536–1268(4), 24–33 (2004)
2. Leonard, J.J., Durrant-Whyte, H.F.: Mobile Robot Localization by Tracking Geometric Beacons. IEEE Transactions on Robotics and Automation 7(3), 376–382 (1991)
3. Cui, S., Xu, X., Zhao, L., Tian, L., Yang, G.: Research on Mobile Robot's Motion Control and Path Planning. In: Yu, W., He, H., Zhang, N. (eds.) ISNN 2009, Part III. LNCS, vol. 5553, pp. 197–206. Springer, Heidelberg (2009)
4. Gracanin, D.: A service-centric model for wireless sensor networks. IEEE Journal on Selected Areas in Communications 23(6), 1159–1166 (2005)

5. Batalin, M.A., Sukhatme, G.S.: The Design and Analysis of an Efficient Local Algorithm for Coverage and Exploration Based on Sensor Network Deployment. IEEE Transaction on Robotics 23(4), 661–675 (2007)
6. Habib, M.K.: Real Time Mapping and Dynamic Navigation for Mobile Robots. International Journal of Advanced Robotic Systems 4(3), 323–338 (2007)
7. Liang, H.W., Chen, W.M., Li, S.: Navigation Algorithm for Mobile Object Based on Wireless Sensor Networks. Chinese Journal of Sensors and Actuators 20(7), 1620–1624 (2007)
8. Buragohain, C., Agrawa, D., Suri, S.: Distributed Navigation Algorithms for Sensor Networks. In: 25th IEEE Intl. Conf. on Computer Communications(INFOCOM), Santa Barbara, pp. 1–10 (2006)
9. Li, Q., Rosa, M.D., Rus, D.: Distributed Algorithms for Guiding Navigation Across a Sensor Network. In: 9th Annual Intl. Conf. on Mobile Computing and Networking, San Diego, pp. 313–325 (2003)
10. Liu, Z., Ding, M.L., Wang, Q.: Implementation of WSN Multi-Node Decision Information Fusion in Autonomous Navigation of Robot. ACTA Electronica Sinica 36(12), 2299–2308 (2008)
11. Djekoune, A.O., Achour, K., Toumi, R.: A Sensor Based Navigation Algorithm for a Mobile Robot using the DVFF Approach. International Journal of Advanced Robotic Systems 6(2), 97–108 (2009)
12. Batalin, M.A., Sukhatme, G.S., Hattig, M.: Mobile Robot Navigation Using a Sensor Network. In: IEEE Intl. Conf. on Robotics and Automation, New Orleans, pp. 636–642 (2004)
13. Moon, T.K., Kuc, T.Y.: An Integrated Intelligent Control Architecture for Mobile Robot Navigation within Sensor Network Environment. In: IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems, Sendal, pp. 565–570 (2004)
14. Zeiger, F., Kraemer, N., Schilling, K.: Commanding mobile robots via wireless ad-hoc networks: A comparison of four ad-hoc routing protocol implementations. In: IEEE International Conference on Robotics and Automation, Pasadena, pp. 590–595 (2008)
15. Lee, H., Yang, S.Y., Sung, H.C.: Robot Path Routing for Shortest Moving Distance in Wireless Robotic Sensor Networks. IEICE Trans. on Communications E92-B(11), 3495–3498 (2010)
16. Koenig, S., Likhachev, M.: Fast Replanning for Navigation in Unknown Terrain. IEEE Transaction on Robotics 21(3), 354–363 (2005)

iMac: Strategy-Proof Incentive Mechanism for Mobile Crowdsourcing

Zhenni Feng[2], Yanmin Zhu[2,1], and Lionel M. Ni[3]

[1] Shanghai Key Lab of Scalable Computing and Systems
[2] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[3] Hong Kong University of Science and Technology
{zhennifeng,yzhu}@sjtu.edu.cn, ni@cse.ust.hk

Abstract. *Mobile crowdsourcing* with smartphones advocates the cooperative effort of mobile smartphones to perform a joint distributed sensing task, which has gained growing importance for its potential to support a wide spectrum of large-scale sensing applications. Smartphone users in the real world are *strategic and rational*. Thus, one crucial problem in mobile crowdsourcing with smartphones is to stimulate cooperation from smartphone users. Several major challenges should be addressed. *First*, the actual cost incurred for a sensing task is *private* information and unknown to other users and the mobile crowdsourcing platform. *Second*, smartphone users are strategic, which suggest a user may deliberately misreport its cost (different from the real cost) in order to maximize its own utility. In this paper, we propose a *strategy-proof* incentive mechanism called *iMac* based on the Vickrey-Clarke-Groves (VCG) mechanism. The main idea of *iMac* is to stimulate smartphone users to truthfully disclose their real costs in spite of strategic behavior of the users. *iMac* introduces two main components. The first component determines the allocation of a sensing task to smartphone users given the user costs. And the second component decides the payment to each user. We prove that *iMac* can successfully produce a unique Nash equilibrium at which each user truthfully discloses the cost. Meanwhile, the minimization of the social cost is achieved. Simulation results demonstrate *iMac* achieves the desired design objectives and the overpayment is modest.

Keywords: Strategy-Proof, Mobile Crowdsourcing, Incentive Mechanism.

1 Introduction

Smartphones have almost increasingly became an indefensible device for people's daily life. The multi-functional property realized by various sensors (e.g., camera, 3D accelerometer, gyroscope, and compass) embedded in the smartphones foster various sensing applications [1].

Mobile crowdsourcing [2] [3] advocates the cooperative effort of mobile smartphones to perform a joint distributed sensing task, which has gained growing

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 337–350, 2013.

importance. With the cellular data channel, a smartphone can share its sensed data with other users. By combining the sensed data from numerous smartphones that are geo-distributed over a vast area, it becomes possible to realize large-scale sensing and monitoring applications, such as places characterization [4] and indoor localization [5].

As a motivating example, we consider the scenario that a user called *sensing requester* wants to learn the air pollution distribution of the city in which the user currently stays. Suppose there is a *mobile crowdsourcing platform* on which smartphone users can share their sensed data. Then, the sensing requester can simply post a sensing task for collecting air pollution data on the platform which then advertises such request to smartphone users. Those smartphone users or *sensing smartphones* respond with their sensed data to the requester. By processing the collected air pollution data from smartphones distributed over different places in the city, the requester is able to have a good understanding of the air pollution distribution in the city.

Mobile crowdsourcing is promising for several reasons. *Firstly*, the geographical distributed nature helps to easily collect distributed sensing data, especially those globally distributed sensing. *Secondly*, embedded sensors of smartphones are actually idle most of the time, and mobile crowdsourcing provides a unique opportunity to better utilize these idle resources. *Thirdly*, the number of smartphones that can contribute to a sensing task is large and this effectively guarantees the sensing quality in terms of spatial distribution and necessary redundancy.

One crucial problem of a mobile crowdsourcing system is how to stimulate cooperation from smartphone users. Contributing to a crowdsourcing sensing task, a smartphone should drive its sensors to collect sensing data. This incurs considerable cost of power, computation and bandwidth. In some cases, a smartphone user may physically move to some specific locations in order to collect the required sensing data. As a result, a smartphone user has to spend considerable cost to accomplish a sensing task. As is well known, smartphone users in the real world are *strategic and rational.* Namely, each smartphone user would join to complete a crowdsourcing task if the user can receive a reward that can sufficiently cover the cost [2] [6] [7] [8]. As a general practice, a sensing task requester can provide monetary reward to a smartphone user who has made a contribution to the sensing task.

By introducing such monetary incentive to smartphone users, *it then becomes an immediate problem that how the crowdsourcing platform should allocate the sensing task to smartphone users and how the payment should be made.* It is highly desirable that the following objectives can be achieved: 1) for all smartphone users, each of them has a good utility so that each user would continuously contribute to the mobile crowdsourcing in the system; and 2) for the crowdsourcing platform, the overall *social cost*, the sum of all user costs, can be minimized. The minimization of the social cost is important, which infers the health of the system and also indicates a low overall payment of the sensing task requester.

Many recent studies on crowdsourcing focus on building an infrastructure to satisfy the objective of scalability or security [9] [10]. In fact, users are strategic and rational, who need proper incentives before they would participate in the system, contributing to mobile crowdsourcing applications. Although the scheme of paying money to users in [3] has been proposed, it fails to prevent real-world users from cheating.

To achieve the aforementioned objectives in a mobile crowdsourcing system, several major challenges should be addressed. *First*, the actual cost incurred for a sensing task varies for different users. Such cost is *private* information and unknown to other users and the mobile crowdsourcing platform. *Second*, smartphone users are strategic, which suggest a user may deliberately misreport its cost (different from the real cost) in order to maximize its own utility. *Finally*, the minimization of the social cost is difficult since it should be achieved under the condition that all users should first be stimulated to participate in the mobile crowdsourcing.

In this paper, we propose a *strategy-proof* incentive mechanism called *iMac* based on the Vickrey-Clarke-Groves (VCG) mechanism. The main idea of *iMac* is to stimulate smartphone users to truthfully disclose their real costs in spite of strategic behavior of the users. If a user claims a cost different from the real one, the obtained utility of the user would suffer. In order to achieve the truthfulness of cost disclosure, *iMac* introduces two main components. The first component determines the allocation of a sensing task to smartphone users given the user costs. And the second component decides the payment to each user. We prove that *iMac* can successfully produce a unique Nash equilibrium at which each user truthfully discloses the cost. Meanwhile, the minimization of the social cost is achieved. Simulation results demonstrate *iMac* achieves the desired design objectives.

We have made three major technical contributions in the paper.

- We model the mobile crowdsourcing under strategic behavior of rational smartphone users who have private cost information as a game.
- We propose an incentive mechanism called *iMac*, stimulating the most cost-efficient smartphones to contribute to the sensing task. It achieves both minimized social cost and strategy-proofness.
- We have conducted both theoretical analysis and extensive simulations. Both of them verify that *iMac* achieves strategy-proofness and show that the overpayment for inducing truthful cost disclosure is modest.

The remainder of the paper is organized as follows. Section 2 describes the system model and the game model, and presents some preliminaries of solution concepts. In Section 3, we elaborate the design of *iMac*. The following section theoretically analyzes several important properties of *iMac*. Simulation results in Section 5 further demonstrate the advantages of *iMac*. In Section 6 we review recent related work and conclude our paper in Section 7.

2 Model and Preliminaries

In the section, we first describe the system model and the game model, and then discuss some solution concepts in the game theory that will be used later.

2.1 System Model

In this paper, we consider a mobile crowdsourcing system with a large number of smartphone users. We denote the set of smartphone users by $L = \{1, 2, \cdots, n\}$, where n is the number of smartphone users. Each sensing requester submits its task to the platform. The platform resides on the cloud. It accepts sensing requests and recruits smartphone users for each sensing requester. Note that this paper assumes that the platform is trustworthy and stays neutral from all activities among requesters and smartphones.

A sensing requester r creates a sensing task which is characterized by the demand of sensing time, denoted as d. Every smartphone i is able to contribute to the completion of the sensing task by performing an amount of the sensing time x_i, $x_i \leq d$. Note that for the sensing task, each unit of sensing time performed by different users results in the *indiscriminate or identical* contribution to the completion of the sensing task. For each smartphone i, there is a limit γ_i on the sensing time that i can spend on performing this sensing task, *i.e.*, $x_i \leq \gamma_i$, considering the limited resources on the smartphone, such as battery.

A unit cost t_i is incurred for each smartphone user i to perform a unit time of sensing. The unit cost information of each user is private and unknown to the requester and other users. Every user is strategic, so it may claim a unit cost c_i that is different from the real unit cost t_i, *i.e.*, $c_i \neq t_i$. We call c_i the *claimed cost.*

For each sensing request d, the platform selects a subset W of smartphones and decides the amount of sensing time x_i for each selected smartphone $i \in W$. To complete the sensing task, it is clear that $\sum_{i \in W} x_i = d$. After the task is completed, the requester pays an amount of money p_i to each smartphone user $i \in W$.

The interaction between the requester and the smartphones shown in Fig. 1 is described as follows.

1. The requester r submits a sensing request with sensing time demand d to the platform, which then advertises it to all smartphones.
2. Each smartphone i responds to the platform by sending its claimed cost c_i and capability γ_i.
3. The platform selects a subset of smartphones $W \subseteq L$ to perform the sensing task. It determines the sensing time $\overrightarrow{x} = (x_1, x_2, \cdots, x_n)$ and the payment $\overrightarrow{p} = (p_1, p_2, \cdots, p_n)$ to each smartphone $i \in W$.
4. The platform notifies each smartphone in W of its sensing time x_i and payment p_i.
5. Each smartphone in subset W performs sensing for a period of x_i and sends the sensed data to the requester.
6. The requester pays each smartphone according to $\overrightarrow{p}$.

In the system there can be many sensing requests. For simplification, however, we assume that the sensing requests do not compete with each other. This indicates that a smartphone can serve more than one sensing requests independently and need not trade off among these requests.

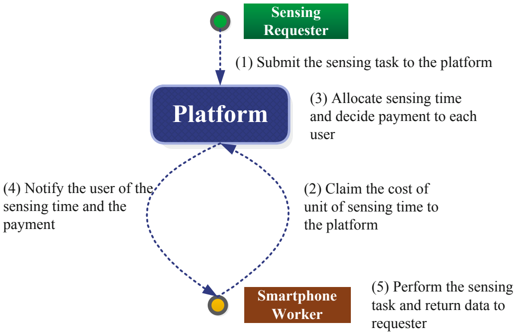

Sensing Requester
(1) Submit the sensing task to the platform
Platform
(3) Allocate sensing time and decide payment to each user
(4) Notify the user of the sensing time and the payment
(2) Claim the cost of unit of sensing time to the platform
Smartphone Worker
(5) Perform the sensing task and return data to requester

Fig. 1. The process of completing a task. A sensing requester submits its sensing time demand d to the platform. Once receiving the sensing request, the platform notice all smartphone users registered in the platform. Each smartphone user i replies the platform with a message containing its claimed cost c_i and maximum length of sensing time γ_i. Then, the platform determined the amount of sensing time for each smartphone user and its corresponding payment. Finally, the smartphone user complete its sensing task and the sensing requester pays to each involved smartphone users.

2.2 Game Model

The interaction among all the smartphones can be modeled as a mathematical game which is characterized by *a set of players*, *a set of actions (strategies)* available to each player, and *a utility vector* corresponding to each combination of strategies. The set of players L consists of all smartphones. The action of each smartphone is to claim the unit cost c_i. The utility u_i of each smartphone i is defined as follows,

$$u_i = \begin{cases} p_i - t_i x_i & i \in W \\ 0 & i \notin W \end{cases}.$$

A smartphone user is willing to participate in the mobile crowdsourcing if its utility is non-negative.

The mobile crowdsourcing platform tries to minimize the social cost which is defined as

$$min \quad \omega = \sum_{j \in W} t_j x_j. \tag{1}$$

The minimization of the social cost is of great importance as it also indicates a modest overall payment of the sensing task requester. Since smartphones are rational, they tend to selfishly report $c_i \neq t_i$ in order to achieve larger utilities. Their strategic behavior contradicts the objective of minimizing the social cost. It is highly desirable that the outcome of the game results in the minimization of the social cost.

2.3 Solution Concepts

In the subsection, we review several important solution concepts in game theory that will be used later.

Definition 1. (Individual-Rationality). *A* mechanism *is individual-rational if each players' utility is non-negative,* i.e., *for each player k, $u_k \geq 0$.*

The definition explains the incentive of rational players to serve for others. A smartphone is willing to participate in the mobile crowdsourcing system if its individual rationality is satisfied.

Definition 2. (Incentive-Compatibility). *A mechanism is incentive-compatible if no player could increase its utility by misreporting its private information (usually referred to as* type*).*

Incentive-compatibility provides an access to disclose the real cost of every smartphone incurred by contributing to the sensing task. On the basis of incentive-compatibility, the mobile crowdsourcing platform is possible to find out the set of most cost-efficient smartphones.

Definition 3. (Strategy-Proof Mechanism). *A mechanism is strategy-proof if it has the property of both individual-rationality and incentive-compatibility.*

Next, we introduce a traditional incentive compatible mechanism called VCG proposed by Vickrey, Clarke and Groves [11]. The VCG mechanism focuses on maximizing the social welfare in a *sealed-bid auction model.* In such an auction model, there are multiple commodities to be sold and many bidders compete to buy commodities. There are many different possible allocations of commodities to bidders, and let the set A denote the set of all allocations. Bidders usually have different preferences upon different allocations $a_k \in A$ and the preference is described by a valuation function v_i for bidder i. The valuation functions are private and unknown to others. The social welfare is defined as $\sum_i v_i(a_k)$ for $a_k \in A$. Each bidder's utility is $u_i(a_k) = v_i(a_k) - p_i(a_k)$, and $p_i(a_k)$ is the price bidder i should pay when allocation of commodities is a_k. The central issue of the mechanism design is to determine price p_i that bidder i should pay so that all players truthfully report their valuations.

3 Design of *iMac*

We propose a strategy-proof mechanism called *iMac* based on VCG mechanism. Its main components are discussed in detail in this section.

3.1 Basic Idea

In our game model, two key issues should be addressed: 1) which smartphones should be chosen to jointly perform the sensing task (the selected smartphones are called *winners*), and 2) how much the requester should pay to each of the involved smartphones.

To address the two issues, *iMac* is designed, which is a mechanism $(g, \overrightarrow{p})$ that satisfies the following two conditions:

- g is an algorithm that determines the optimal sensing time allocation that minimizes the social cost. In another word, g is a mapping that $\overrightarrow{x^*} = g(t_1, t_2, \cdots, t_n)$ where $\overrightarrow{x^*}$ minimizes the social cost.
- $\overrightarrow{p}$ is a payment scheme that determines the payment smartphone i receives from the requester, $p_i = h(t_{-i}) - \sum_{j \in L, j \neq i} x_j^* t_j$, where $h(t_{-i})$ is a function independent of the cost claimed by smartphone i.

Next, we discuss how to determine g and $\overrightarrow{p}$ of *iMac* in detail. To simplify the discussion, we first assume that $c_i = t_i$ for each i, indicating that each smartphone disclose its real cost honestly. Then, we show that such honest report is an equilibrium so that we can predict each smartphone will not violate the equilibrium.

3.2 Selecting Winners

In the subsection, we first model the winner selection as a linear programming problem and then find its optimal solution. The optimal solution should contain both the set of winners and their sensing times.

The winner selection problem is

$$\begin{aligned} min \ \ \omega = \sum_{i \in L} t_i x_i & \\ 0 \leq x_i \leq \gamma_i, \forall i \in L & \qquad (2) \\ \sum_{i \in L} x_i = d. & \qquad (3) \end{aligned}$$

Here the objective has been discussed before. (2) tells that the sensing time of each smartphone i could not exceed its capability γ_i while $x_i = 0$ means smartphone i is not chosen. Then, (3) shows that the demand of the requester must be satisfied.

Next, we propose a greedy algorithm shown in Algorithm 1 to determine the sensing time of each smartphone and compute the social cost of completing the sensing task.

3.3 Determining Payment

After choosing the set of winners, we need to determine the payment to each of them. We introduce *the Clarke pivot rule* to help us determine the payment function.

In the previous auction model, bidder i pays an amount equal to the damage of social welfare it causes to all other bidders in the system, *i.e.*, the difference

Algorithm 1. Computing sensing time $(x_1, x_2, \cdots, x_n)$ and social cost ω

```
Input: sensing demand $d$, claimed costs $\overrightarrow{c} = \overrightarrow{t} = (t_1, t_2, \cdots, t_n)$, capabilities $\overrightarrow{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_3)$.
Output: sensing time $\overrightarrow{x} = (x_1, x_2, \cdots, x_3)$, social cost $\omega$.
$\overrightarrow{x} \leftarrow \overrightarrow{0}$
$\omega \leftarrow 0$
$m \leftarrow d$
sort the smartphone in the nondecreasing order and store the order in a list $\mathcal{L}$
while $m > 0$ do
    if length($\mathcal{L}$)is 0 then
        PRINT 'no solution'
        RETURN
    else
        $i$ ←head of $\mathcal{L}$
        delete head of $\mathcal{L}$
        if $m \geq \gamma_i$ then
            $m \leftarrow m - \gamma_i$
            $x_i \leftarrow \gamma_i$
        else
            $m \leftarrow 0$
            $x_i \leftarrow m$
        end if
        $\omega \leftarrow \omega + x_i t_i$
    end if
end while
```

between all other bidders' maximized social welfare with and without bidder i. In our problem, the payment is determined similarly. A smartphone will be paid an amount equal to the benefit it introduces to the system, *i.e.,* the difference between others smartphones' minimized social cost with and without it. The payment scheme is

$$p_i = \omega^*_{-i} - (\omega^* - t_i x^*_i), \forall i \in L. \tag{4}$$

Here ω^*_{-i} means the minimized social cost without smartphone i. Particularly, when a smartphone j is not chosen, it introduces nothing to the system, so it will be paid nothing. And in (4), $\omega^*_{-j} = \omega^*$ and its utility is zero. Then, a smartphone $i's$ utility is

$$u_i = \omega^*_{-i} - \omega^*. \tag{5}$$

The payment computation algorithm is shown in Algorithm 2.

4 Analysis

In the section, we discuss several properties of *iMac* and prove that that it achieves our goals through theoretical analysis.

Lemma 1. *iMac satisfies the property of individual-rationality.*

Algorithm 2. Compute Payment $(p_1, p_2, \cdots, p_n)$

```
Input: sensing demand $d$, claimed costs $\overrightarrow{c} = \overrightarrow{t} = (t_1, t_2, \cdots, t_n)$, capabilities $\overrightarrow{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_3)$.
Output: payment $\overrightarrow{p} = (p_1, p_2, \cdots, p_3)$.
$\overrightarrow{p} \leftarrow \overrightarrow{0}$
compute the minimized social cost $\omega^*$and sensing time $\overrightarrow{x^*} = (x_1^*, x_2^*, \cdots, x_n^*)$
for $i = 1 \rightarrow n$ do
    if $x_i^* > 0$ then
        compute the minimized social cost $\omega_{-i}^*$ without smartphone $i$
        $p_i = \omega_{-i}^* - (\omega^* - t_i x_i^*)$
    end if
end for
```

Proof. We denote the optimal sensing time is $\overrightarrow{x^*} = (x_1^*, x_2^*, \cdots, x_n^*)$ and the optimal social cost is ω^*. For a smartphone i that is not chosen, its utility $u_i = 0$. For a smartphone i belonging to the winner set, its utility is

$$\begin{aligned} u_i &= p_i - x_i^* t_i \\ &= \omega_{-i}^* - (\omega^* - t_i x_i^*) - x_i^* t_i \\ &= \omega_{-i}^* - \omega^* \\ &\geq 0. \end{aligned}$$

Here $\omega_{-i}^* \geq \omega^*$ because ω^* denotes an optimal solution.

Lemma 2. *iMac is incentive-compatible.*

Proof. For a smartphone i, we fix other smartphones' real costs $\overrightarrow{t_{-i}} = (t_1, \cdots, t_{i-1}, t_{i+1}, \cdots, t_n)$ and the ω_{-i}^* will not change. The smartphone i's utility is $u_i = \omega_{-i}^* - \omega^*$. If the smartphone i reports a claimed cost of $c_i = t_i' \neq t_i$, the corresponding optimal sensing time and social cost will be $\overrightarrow{x}'$ and ω'. Consequently, its utility will be $u_i' = \omega_{-i}^* - \omega' = \omega_{-i}^* - \omega^* \leq u_i$ due to $\omega^* \leq \omega'$. To sum up, each smartphone i could not increase its utility by misreporting its private information t_i.

Theorem 1. *iMac is strategy-proof.*

Proof. In Lemma 1 and Lemma 2, we have proved that *iMac* is individual-rational and incentive-compatible.

5 Performance Evaluation

In the section, we evaluate the performance of *iMac* through simulations and the results are discussed.

5.1 Methodology and Simulation Setup

We design a baseline mechanism to compare with *iMac*. The baseline mechanism just ignores the strategic behavior. Consequently, the requester just pays what the smartphones claim. In the baseline mechanism, smartphones are likely to cheat. A rational smartphone will claim a higher cost than its real cost. Nevertheless, it could not claim too much since it has the risk of not being chosen as a winner. To model the misreporting of a smartphone i, we assume that each smartphone i randomly chooses a value from $[t_i, t_i\theta]$ where θ denotes *the misreporting ratio*.

We evaluate the mechanisms with the following metrics:

- **Social cost**: the sum of the real costs of all involved smartphone users for completing the sensing task.
- **Overpayment**: the ratio of the total payment to the social cost.

We assume that the real costs and the capabilities of the smartphones follow two typical models: uniform distribution and normal distribution. For the uniform distribution, a real cost is uniformly distributed in [5, 20] and a capability is uniformly distributed in [20,50]. For normal distribution, the real cost has the mean of $\mu = 10$ and the standard deviation of $\sigma = 3$. In addition, the capability has the mean of $\mu = 30$ and the standard deviation of $\sigma = 8$.

The default system parameters are set as follows. The number of smartphones is $n = 1,000$, the sensing demand is $d = 1,000$, and the misreporting ratio is $\theta = 3$.

5.2 Comparison of Overpayment

In our first set of simulations, we show that the overpayment introduced by *iMac* is much smaller than that of the baseline mechanism except when all smartphones always honestly report their real cost, which is not possible in reality.

In Fig. 2, we can see that the overpayment decreases when more smartphones can be candidates for both mechanisms. This is because when the number of smartphones goes up, there tends to be more cheap smartphones. *iMac* and baseline mechanism combined with uniform distribution or normal distribution share descending tendency. Particularly, the overpayment of *iMac* with uniform distribution or normal distribution is considerably close to one when $n = 2,500$, while the overpayment of the baseline mechanism with uniform distribution or normal distribution is still as high as 1.25 and 1.3, respectively.

In Fig.3, the overpayment of either mechanism with either distribution increases when the task demands longer sensing time. This is because longer sensing time needs more smartphones to participate and thus some more expensive smartphones will be chosen. However, the overpayment of *iMac* with uniform distribution and normal distribution is always lower than 1.1, showing its advantage over the baseline mechanism.

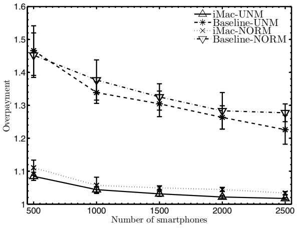


Fig. 2. Overpayment vs. Number of smartphones n. ($d = 1,000, \theta = 3$)

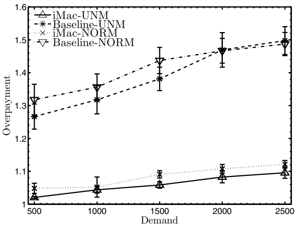


Fig. 3. Overpayment vs. Demand d. ($n = 1,000, \theta = 3$)

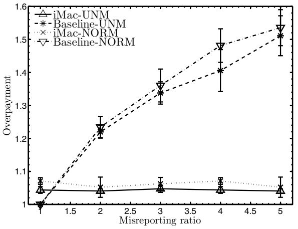


Fig. 4. Overpayment vs. Misreporting ratio θ. ($n = 1,000, d = 1,000$)

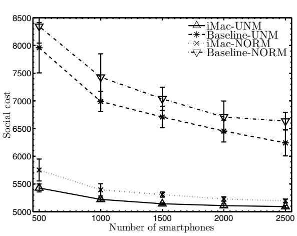


Fig. 5. Social cost vs. Number of smartphones n. ($d = 1,000, \theta = 3$)

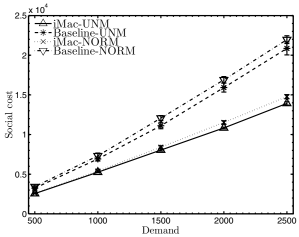


Fig. 6. Social cost vs. Demand d. ($n = 1,000, \theta = 3$)

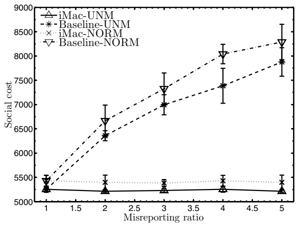


Fig. 7. Social cost vs. Misreporting ratio θ. ($n = 1,000, d = 1,000$)

In Fig. 4, we can see that when smartphones cheat and misreports a cost, overpayment of the baseline mechanism with both distributions rises sharply while *iMac* still keeps a very low level of about 1.05 with uniform distribution or normal distribution. This figure reveals that *iMac* with both uniform distribution and normal distribution is more resilient to strategic behavior than the baseline mechanism. When smartphones are completely honestly ($\theta = 1$), there is no overpayment for the baseline mechanism. In addition, when smartphones charges slightly larger with $\theta = 1.2$, *iMac* with uniform distribution shares the same overpayment with the baseline mechanism using either distributions.

5.3 Comparison of Social Cost

In our second set of simulations, we show that the social cost introduced by *iMac* is always smaller than that of the baseline mechanism for both uniform distribution and normal distribution.

Fig. 5 explains the social cost increases when there are more smartphones willing to participate. Given the fixed demand, there will be more cheap smartphones. Thus, the social costs of the two mechanism share a descending tendency when n rising for either distribution model. From the picture, we can also see that the baseline mechanism consumes at least 30% more social cost than *iMac* for both the uniform distribution and the normal distribution.

As shown in Fig.6, the social costs of both mechanisms experience uptrend for uniform distribution and normal distribution when demand d rises. It is natural that longer sensing demand consumes more resources. We can also know that the baseline mechanism goes up much faster than *iMac*. When the demand d is 2,500, the baseline mechanism consumes about 2×10^4 social cost while that of *iMac* is less than 1.5×10^4 for both distributions.

Fig. 7 shows that it is more possible for *iMac* to select the most cost-efficient smartphones for both uniform and normal distribution. When smartphones always report their real costs honestly ($i.e.,\ \theta = 1$), social cost determined by two mechanisms are the same. Actually, when users become strategic, the social cost of the baseline mechanism experiences a sharp rise while *iMac* still keeps a low level of about 5,250 for both uniform distribution and normal distribution.

6 Related Work

Mobile crowdsourcing is faced up with many key unsolved issues, such as scalability, security. Thus, a lot of papers has been published to study the framework of mobile crowdsourcing especially in the scenario of mobile phone sensing.

Dong *et al.* propose Zoom [9] which focuses on how to assign the tasks to heterogeneous and geographically distributed phones. They divide the numerous phones into groups of arbitrary size and perform the task assignment on the level of groups.

In [10] and [12], the authors design their frameworks considering the context of tasks. In the architecture described in [10], people's preferences for media data are different and then the system tries to assign the right tasks to the right people in the proper circumstances. Jayaraman *et al.* explain several key issues such as scalability, real-time problems in their paper [12] and give a simple demonstration scenario.

Moreover, both [12] and [13] introduce cloud based frameworks into mobile crowdsourcing. To achieve the object of scalability, the authors think out a cloud-assisted crowdsourcing framework [13]. The framework assumes the existence of distributed clouds near applications, thus separating the data collection and sharing. To handle the heterogeneity of smartphone platform, they introduce standard APIs. All the design requirements targets to scale up the crowdsourcing system.

Different from the above papers, the authors propose Medusa and implement the system from the viewpoint of programming framework [3], demonstrating many detailed issues such as privacy, security, incentives, coordination.

In [14], the authors try to support the data collection problem in crowdsourcing application by the methods of data mining. The procedure of data management (mining) helps to reduce the amount of sent data and it is consistent with the target of energy efficiency.

Yang *et al.* discuss the incentive mechanism using two kinds of models [6]. The first model is platform-centric model where the platform has the priority of deciding the payment for every unit of sensing time and it is solved by Stackelberg game [15]. The second model is user-centric model where each user keeps its cost as private information. The platform announces the tasks, and each user could finish a part of the task. However, in the user-centric model, the competition results of users can not guarantee all parts of the task are finished.

7 Conclusion

In this paper, we have focused on the crucial problem of provisioning incentives in mobile crowdsourcing. Due to the strategic behavior and the private cost information of smartphone users , stimulating cooperation of smartphone users is particularly challenging. In this paper we have designed *iMac* which is a strategy-proof mechanism based on the VCG mechanism. *iMac* stimulates the smartphones to honestly report their real costs, and then achieves the minimized social cost. The theoretical analysis and the simulation results verify that it is strategy-proof, *i.e.*, all smartphones are willing to participate in mobile crowdsourcing in spite of their selfishness. And importantly, the sensing requester needs to pay only a modest overpayment for realizing the truthfulness of real cost disclosure.

Acknowledgements. This research is supported by NSFC (No. 61170238, 60903190, 61027009, 60933011, 61202375, 61170237), Shanghai Pu Jiang Talents Program (10PJ1405800), Shanghai Chen Guang Program (10CG11), MIIT of China (2009ZX03006-001-01), Doctoral Fund of Ministry of Education of China (20100073120021), National 863 Program (2009AA012201 and 2011AA010500), HP IRP (CW267311), SJTU SMC Project (201120), STCSM (08dz1501600, 12ZR1414900), Singapore NRF (CREATE E2S2), and Program for Changjiang Scholars and Innovative Research Team in Universities of China (IRT1158, PCSIRT).

References

1. Nath, S.: Ace: exploiting correlation for energy-efficient and continuous context sensing. In: Proc. ACM MobiSys. (2012)
2. Ganti, R., Ye, F., Lei, H.: Mobile crowdsensing: current state and future challenges. IEEE Communications Magazine 49(11), 32–39 (2011)
3. Ra, M.R., Liu, B., La Porta, T.F., Govindan, R.: Medusa: a programming framework for crowd-sensing applications. In: Proc. ACM MobiSys. (2012)
4. Chon, Y., Lane, N.D., Li, F., Cha, H., Zhao, F.: Automatically characterizing places with opportunistic crowdsensing using smartphones. In: Proc. of the 2012 ACM Conference on Ubiquitous Computing (2012)
5. Rai, A., Chintalapudi, K.K., Padmanabhan, V.N., Sen, R.: Zee: zero-effort crowdsourcing for indoor localization. In: Proc. ACM MOBICOM (2012)
6. Yang, D., Xue, G., Fang, X., Tang, J.: Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing. In: Proc. ACM MOBICOM (2012)
7. Anderegg, L., Eidenbenz, S.: Ad hoc-vcg: A truthful and cost-efficient routing protocol for mobile ad hoc networks with selfish agents (2003)

8. Wang, W., Li, X.Y., Wang, Y.: Truthful multicast routing in selfish wireless networks. In: Proc. ACM MOBICOM (2004)
9. Dang, T., Chi Feng, W., Bulusu, N.: Zoom: A multi-resolution tasking framework for crowdsourced geo-spatial sensing. In: Proc. IEEE INFOCOM (2011)
10. Tamilin, A., Carreras, I., Ssebaggala, E., Opira, A., Conci, N.: Context-aware mobile crowdsourcing. In: Proc. of the 2012 ACM Conference on Ubiquitous Computing (2012)
11. Nisan, N., Roughgarden, T., Tardos, E., Vazirani, V.V.: Algorithmic game theory. Cambridge University Press (2007)
12. Jayaraman, P., Sinha, A., Sherchan, W., Krishnaswamy, S., Zaslavsky, A., Haghighi, P.D., Loke, S., Do, M.T.: Here-n-now: A framework for context-aware mobile crowdsensing. In: Proc. of the Tenth International Conference on Pervasive Computing (2012)
13. Xiao, Y., Simoens, P., Pillai, P., Ha, K., Satyanarayanan, M.: Lowering the barriers to large-scale mobile crowdsensing. In: Proc. of the 14th Workshop on Mobile Computing Systems and Applications (2013)
14. Sherchan, W., Jayaraman, P.P., Krishnaswamy, S., Zaslavsky, A., Loke, S., Sinha, A.: Using on-the-move mining for mobile crowdsensing. In: Proc. of the 13th IEEE International Conference on Mobile Data Management, MDM 2012 (2012)
15. Fudenberg, D., Tirole, J.: Game theory. MIT Press (1991)

Social Welfare Maximization in Participatory Smartphone Sensing

Tong Liu[1] and Yanmin Zhu[1,2]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] Shanghai Key Lab of Scalable Computing and Systems
{liutong25691640,yzhu}@sjtu.edu.cn

Abstract. *Participatory smartphone sensing* has lately become more and more popular as a new paradigm for performing large-scale sensing, in which each smartphone contributes its sensed data for a collaborative sensing application. Most existing studies assume that smartphone users are strictly strategic and completely rational, which can achieve only sub-optimal system performance. Few existing studies can maximize a system-wide objective which takes both the platform and smartphone users into account. This paper focuses on the crucial problem of maximizing the system-wide performance or social welfare for a participatory smartphone sensing system. There are two great challenges. *First*, the social welfare maximization can not be realized on the platform side because the cost of each user is *private* and *unknown* to the platform in reality. *Second*, the participatory sensing system is a large-scale real-time system due to the huge number of smartphone users who are geo-distributed in the whole world. We propose *a novel price-based decomposition framework*, in which the platform provides a unit price for the sensing time spent by each user and the users return the sensing time via maximizing the monetary reward. This pricing framework is an effective incentive mechanism as users are motivated to participate for monetary rewards from the platform. The original problem is equivalently converted into an optimal pricing problem, and a distributed solution via a *step-size-free* price-updating algorithm is proposed. More importantly, the distributed algorithm ensures that the cost privacy of each user is not compromised. Experimental results show that our novel distributed algorithm can achieve the maximum social welfare of the participatory smartphone system.

Keywords: Participatory smartphone sensing, pricing, distributed optimizations.

1 Introduction

Because of the fast progress of wireless communication technology, such as the third-generation mobile telecommunications (3G), the proliferation of smartphones has been witnessed in the past few years. Smartphones are increasingly equipped with a set of cheap but useful sensors such as accelerometer, camera, and GPS, which can collect a lot of diverse useful information of our daily life.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 351–365, 2013.

Participatory sensing [1] has lately become more and more popular as a new paradigm for performing large-scale sensing, in which each node contributes its sensed data for a collaborative sensing application. Smartphones are usually distributed over the globe and can stay connected over the Internet. This leads to the construction of *participatory smartphone sensing* [2] [3]. A lot of appealing applications can be realized, such as noise mapping and environment quality monitoring, which are almost impossible in the past. Smartphones essentially constitute a large-scale system of mobile sensors, instead of traditional static sensors. A participatory smartphone sensing system consists of a sensing *platform* and many smartphone *users*, which connect with the platform via cellular data channels. In such a system of participatory smartphone sensing, the platform first advertises a task of smartphone sensing, and smartphone users registered with the system can participate in such a task by sending the sensed data to the platform.

Because of the potential advantage of participatory smartphone sensing, a great deal of research has devoted to developing various applications [4] [5] [6], systems [7] [8] and technologies [9]. These existing studies simply assume that all users are bounded to be cooperative and follow design instructions. In the real-world, however, smartphones consume resources such as power and time. Therefore, incentive mechanisms are needed to stimulate them to participate distributed sensing. As a general incentive mechanism, users get monetary reward for providing sensed data.

Many existing studies [10] [11] [12] [13] [14] employ *auction* algorithms to provide incentive to users. These existing studies usually assume that smartphone users are strictly strategic and completely rational. As a result of such studies, Nash equilibria are computed at which each player will not change its strategy unilaterally if other players do not change their strategies. The Nash equilibria usually lead to sub-optimal system performance. Thus, none of such studies can maximize a system-wide objective which takes both the platform and users into account. The concept of *social welfare* is important, which is usually used to describe some system-wide performance objectives.

This paper focuses on maximizing the system-wide performance or social welfare for a participatory smartphone sensing system. More specifically, the social welfare is defined as the platform's utility of accomplishing a sensing task minus the total costs of smartphone users. However, there are two great challenges for achieving such a system-wide performance objective. *First*, the social welfare maximization cannot be realized on the platform side because the cost of each user is *private* and *unknown* to the platform in reality. *Second*, the participatory sensing system is a large-scale real-time system due to the huge number of smartphone users who are geo-distributed in the whole world. Note that the optimal objective is coupled, which cannot be decomposed into several subproblems respectively achieved by the platform and each user.

To overcome the difficulties, we propose a novel *price-based decomposition framework*, where alternatively the platform provides a unit price for the sensing time spent by each user and the users return the selection of sensing time via maximizing the monetary reward minus the cost. This pricing framework is a kind of an incentive mechanism as users are motivated to participate in the participatory sensing for getting monetary rewards from the platform. The original problem is converted into an equivalent optimal pricing problem, and a distributed solution via a *step-size-free* price-updating algorithm is proposed.

The main technical contributions of this paper are summarized as follows.

- We formally formulate the problem of maximizing the social welfare of a participatory smartphone sensing system, which is difficult to be solved due to its coupled objective.
- We propose an incentive mechanism for maximizing the social welfare. First, the pricing framework is developed, exploiting the correlation between the prices provided by the platform and the selected sensing time by users. Next, we propose a *step-size-free* algorithm for price updating and formally prove the convergence condition. The main feature of the distributed algorithm is that it ensures that the cost privacy of each user is not compromised.
- We have performed extensive simulations. Experimental results show that our novel distributed algorithm can successfully maximize the social welfare of the system.

The rest of the paper is organized as follows. We review related work in Section 2. In Section 3, we describe the system model and formulate the problem of social welfare maximization. The problem is converted to an equivalent pricing problem in Section 4. Section 5 presents the details of the distributed algorithm for computing the optimal price. We evaluate the proposed algorithm via simulations in Section 6 and conclude in Section 7.

2 Background and Related Work

Participatory smartphone sensing has attracted a lot of researches. Several systems and applications have been developed, such as mCloud [7] which is an iPhone-based mobile crowdsourcing platform, S^2aaS [8] to provide sensing services for could users, PIER [4] for reporting personal environmental impact and NoiseTube [5] for measuring and mapping the urban noise pollution. However, most of these systems and applications focus on the system design by assuming there are numerous smartphones voluntarily participating in the system. Actually, smartphone users are motivated to participate by some money or service reward because of resources consuming for sensed data, such as power and time.

Several recent works have focused on designing incentive mechanisms to motivate smartphone users for participatory sensing. Most of these works apply *auction* algorithms. A novel reverse auction based on dynamic price is proposed in [10] [11], where users sell their sensed data to the platform by claiming bid prices, and the platform selects a subset of these users and purchases at their bid prices. Their work shows that using dynamic price can reduce the incentive cost compared with fixed price. Luis G. Jaimes et al. [12] propose a greedy algorithm based on the recurrent reverse auction incentive mechanism, which aims to maximize the covered area under a budget constraint. Two system models are considered in [13]: the platform-centric model and the user-centric model. For the platform-centric model, an incentive mechanism based on Stackelberg game is designed to maximize the utility of the platform. For the user-centric model, an auction-based incentive mechanism is designed, under individuals rational assumption. Koutsopoulos et al. [14] aim to minimize the total costs of compensating users via proposing an optimal reverse auction mechanism. This work illustrates the significance of the assumptions that the costs of users are their privacy, and users are strategic to maximize their own utilities.

None of all the above works applying incentive mechanisms try to maximize the social welfare of a sensing system. Here, *Social welfare* is a system-wide performance metric, and is defined as the total surplus of the sensing system, which is the utility of accomplished sensing tasks minus the costs spent by users. From the viewpoint of the system, maximizing the social welfare is naturally desired. To the best of our knowledge, our work is the first attempt which considers the design of an incentive mechanism for maximizing the system-wide social welfare under the condition that the costs of users are unknown to the platform.

3 System Model and Problem Formulation

3.1 System Model

The participatory sensing system consists of a *platform*, which resides in the cloud, and a large number of *smartphone users*, which are connected to the platform via cellular data channels. In such a system, the platform has a sensing task of collecting a large amount of distributed sensing data and there is a set of users, denoted as $\mathcal{U} = \{1, 2, \cdots, n\}$, who are interested in contributing to the sensing task, where $n \geq 2$. Each user participating in the task will sense for a period of time and then send the sensed data to the platform. The platform collects the sensed data from users to extract valuable information.

For user $i \in \mathcal{U} = \{1, 2, \cdots, n\}$ who participates in the task, it incurs a *cost*, which is a function of the time t_i spent on the task. We assume that each user promises the amount of sensing time W_i after knowing the description of the platform's sensing demand, which is specific for different user i. The sensing time W_i can be considered as a constant given the specific demand, which is only decided by the practical situation of user i and the specific demand. We model the cost $C_i(t_i)$ of user i as a function of both actual sensing time t_i and the demand W_i. For example, a user may have a linear cost $\alpha_i t_i$ for the actual sensing time t_i and an extra convexly increasing cost $e^{\beta_i(t_i - W_i)}$ for the time beyond the promised sensing time, where α_i, β_i are user-specific parameters. We first define a notation $\phi_i(t_i)$ for the cost beyond the promised sensing time as follows,

$$\phi_i(t_i) = \begin{cases} 0 & \text{w.p. } t_i \leq W_i \\ e^{\beta_i(t_i - W_i)} & \text{w.p. } t_i > W_i \end{cases} . \tag{1}$$

Therefore, the total cost $C_i(t_i)$ of user i is defined as follows.

Definition 1 (User Cost). *The cost $C_i(t_i)$ of user i is the sum of the linear incremental cost of sensing time and the potential cost due to unwilling extra time, which is formulated as*

$$C_i(t_i) = \alpha_i t_i + \phi_i(t_i). \tag{2}$$

Note that C_i is *convex* and *monotonically increasing* with t_i.

Due to the cost, each user participating in the sensing expects a *payment* in return from the platform. We assume that the platform provides a unit price p_i for the time spent by user i as the reward. We define the *payoff* of user i in the following.

Definition 2 (User Payoff). *The* payoff *of user i is the difference between the payment and cost, which is formulated as*

$$P_i(t_i, p_i) = p_i t_i - C_i(t_i). \tag{3}$$

Given price p_i, a rational user i will choose a t_i to maximize its payoff P_i.

With collecting sensed data and extracting useful information, the platform can obtain a certain quantity of *utility* gain, which is related to the actual sensing time $\mathbf{t} = [t_1, \cdots, t_n]$ spent by all users. We give the formulation of the utility in the following.

Definition 3 (Platform Utility). *The utility of the platform is positive related with each user's sensing time, i.e.,*

$$U(\mathbf{t}) = \lambda \log(1 + \sum_{i \in \mathcal{U}} \log(1 + t_i)), \tag{4}$$

where λ is a system parameter.

A *log* function used here is according to the law of diminishing marginal utility in economics. The $\log(1 + t_i)$ term reflects the diminishing utility gain on the sensing time of user i, and the outer log term reflects the diminishing utility gain on the number of participating users. Such an expression has been adopted for the crowdsourcing system in other literature [13].

As mentioned above, the platform should pay for the sensing time spent by each user at price p_i. We define the *profit* of the platform as the difference between its utility and its payment.

Definition 4 (Platform Profit). *The profit of the platform is formulated as*

$$\Psi(\mathbf{t}, \mathbf{p}) = U(\mathbf{t}) - \mathbf{p}^T \mathbf{t}. \tag{5}$$

3.2 Problem Formulation

In this paper, we study the sensing platform as a social organization whose objective is to maximize *social welfare* $S(\mathbf{t})$, which is defined as follows.

Definition 5 (Social Welfare). *The social welfare of the system is the platform's utility minus the total costs of all users,*

$$S(\mathbf{t}) = U(\mathbf{t}) - \sum_{i \in \mathcal{U}} C_i(t_i). \tag{6}$$

The platform aims to decide the optimal sensing time $\mathbf{t}^* = [t_1^*, \cdots, t_n^*]^\mathsf{T}$ for users by solving the problem of maximizing the system's social welfare, which is defined as follows.

Definition 6 (Social Welfare Maximization Problem). *The problem of maximizing the social welfare of the participatory smartphone sensing system is defined as,*

$$\begin{aligned} \max_{\mathbf{t}} \quad & U(\mathbf{t}) - \sum_{i \in \mathcal{U}} C_i(t_i) \\ s.t. \quad & t_i \in [0, \tau_i], \forall i \in [1, n], \end{aligned} \tag{7}$$

where τ_i is the upper bound of time t_i, which can be infinite.

Note that $U(\mathbf{t})$ is a strictly concave function that is monotonically increasing in each t_i. Therefore, the problem (7) is a convex optimization problem. There exists a unique solution $\mathbf{t}^* = [t_i^*, \cdots, t_n^*]$ which represents the best sensing time for each user.

This problem can be seen as a resource allocation problem in a distributed system. Due to the large number of users, the optimization problem becomes difficult to solve and a distributed solution is necessary which avoids the single-point-of-failure problem as well.

4 Pricing for Social Welfare Maximization

The distributed solution to problem (7) can be achieved via designing an iterative pricing framework. In this section, we first introduce the pricing framework. Based on the pricing framework, problem (7) can be converted to a pricing problem.

4.1 Pricing Framework

As mentioned in Section 3.1, the platform pays for the sensing services provided by users by setting a price vector $\mathbf{p} = [p_1, \cdots, p_n]$. When a price p_i is given by the platform, a rational user will choose a sensing time t_i to maximize the payoff.

$$\widetilde{t_i}(p_i) = \underset{t_i \in [0, \tau_i]}{\arg\max} \quad P_i(t_i, p_i) = p_i t_i - C_i(t_i). \tag{8}$$

Each $\widetilde{t_i} = \widetilde{t_i}(p_i)$ is computed *locally* by user i and sent to the platform. Based on all the returned $\widetilde{t_i}$, the provider will *update* the price vector $\mathbf{p}$ iteratively with

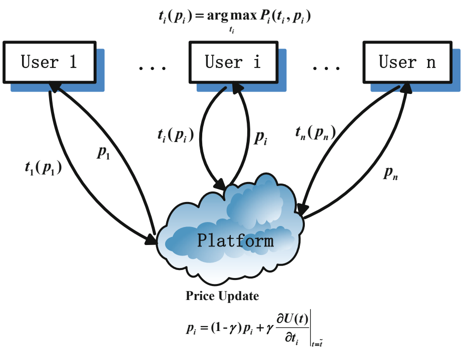

$t_i(p_i) = \arg\max_{t_i} P_i(t_i, p_i)$
User 1
...
User i
...
User n
$t_1(p_1)$
p_1
$t_i(p_i)$
p_i
$t_n(p_n)$
p_n
Platform
Price Update
$p_i = (1-\gamma)p_i + \gamma \frac{\partial U(t)}{\partial t_i}\Big|_{t=\tilde{t}}$

Fig. 1. The illustration of pricing framework

the objective that the $\widetilde{\mathbf{t}}(\mathbf{p})$ produced by maximizing users' payoff will eventually converge to the optimal solution $\mathbf{t}^*$. Such a pricing framework is interpreted in Fig. 1.

In this pricing framework, updating prices does not need to be executed after each $\widetilde{t}_i$ is returned. The platform can update p_i for user i for arbitrary times before $p_j (j \neq i)$ is updated. That is to say, our pricing framework can be realized in an asynchronous way, such that the time of convergence can be largely degraded as different user has different message-transmitting delay.

To apply this pricing framework in a sensing system, the platform and all users should sign an agreement on a pricing process before users begin participate in the task. *In the pricing process, the platform first provides an initial price and each user returns the sensing time via maximizing the payoff. The price vector is updated iteratively by the platform until the social welfare is maximized given the sensing time returned by all users.* As a result, the price vector and sensing time vector are determined and all users begin to sense under this agreement.

4.2 Converting to Optimal Pricing Problem

Given a price vector $\mathbf{p} = [p_1, \cdots, p_n]$, the social welfare can be rewritten as the sum of the platform's profit and the payoffs of users as follows,

$$\begin{aligned} S(\mathbf{t}) &= (U(\mathbf{t}) - \mathbf{p}^T\mathbf{t}) + \sum_{i \in \mathcal{U}} (p_i t_i - C_i(t_i)) \\ &= \Psi(\mathbf{t}, \mathbf{p}) + \sum_{i \in \mathcal{U}} P_i(t_i, p_i). \end{aligned} \tag{9}$$

Furthermore, a rational user will maximize its own payoff given a price as in (8), such that the social welfare maximization problem (7) can be converted into an equivalent **optimal pricing problem**.

Definition 7 (Optimal Pricing Problem). *Given a price p_i, user i will choose a sensing time $\widetilde{t_i}$ to maximize its payoff, i.e., $\widetilde{t_i}(p_i) = \arg\max_{t_i \in [0,\tau_i]} p_i t_i - C_i(t_i)$. The optimal pricing problem is to determine the prices $p_1, \cdots, p_n$ by maximizing the social welfare,*

$$\max_{\mathbf{p}} \quad \Psi(\widetilde{\mathbf{t}}, \mathbf{p}) + \sum_{i \in \mathcal{U}} P_i(\widetilde{t_i}, p_i). \tag{10}$$

Assume the optimal solution of problem (10) is $\mathbf{p}^*$. Now we illustrate why the optimal pricing problem (10) is equivalent to the original problem (7). When the optimal price p_i^* is given to user i, $\widetilde{t_i}(p_i^*)$ is chosen by maximizing the payoff (8). Given $\widetilde{t_i}(p_i^*)$ and $\mathbf{p}^*$, the platform's profit is maximized, such that the social welfare maximization is achieved according to (9). In other words, $\widetilde{t_i}(p_i^*)$ is exactly the optimal sensing time t_i^* for user i. Therefore, we just need to find the optimal prices $p_1^*, \cdots, p_n^*$ based on the iterative pricing framework in Section 4.1.

5 Distributed Payment Solution

5.1 Overview

As we have converted the original social welfare maximization problem in (7) to the optimal pricing problem in (10), the main challenge now is to design a price updating rule for the platform, so that the sensing time computed by users can quickly converge to the optimal value. We propose a new price updating rule as follows,

$$p_i = (1-\gamma)p_i + \gamma \left.\frac{\partial U(\mathbf{t})}{\partial t_i}\right|_{\mathbf{t}=\widetilde{\mathbf{t}}}, \quad \forall i, \tag{11}$$

where $\gamma \in (0, 1]$ is a tunable relaxation parameter. After receiving a return $\widetilde{t_i}$ from user i, the platform generates a new price based on the original price and the partial derivative $\left.\frac{\partial U(\mathbf{t})}{\partial t_i}\right|_{\mathbf{t}=\widetilde{\mathbf{t}}}$, which pushes the price p_i towards to the optimum $p_i^* = \frac{\partial U(\mathbf{t}^*)}{\partial t_i}$.

5.2 Distributed Algorithm

The distributed algorithm given in Algorithm 1 is based on the alternation of two processes:

- **Process 1**: the price updating according to the rule in (11) via the platform,
- **Process 2**: the payoff maximizing according to (8) via users.

At the beginning of Algorithm 1, the platform sets the initial price $p^{(0)}$ for $\forall i$, which is related to the convergence of the algorithm. According to the convergence condition given in Section 5.3, $p^{(0)}$ can be set to an arbitrary value when $\gamma = 1$ and be set extremely close to zero when $\gamma < 1$. Uniformly, we set $p^{(0)}$ to a very little positive constant ϵ. Then, each user updates the sensing time $\widetilde{t_i}$ via (8) and the platform updates the prices after receiving $\widetilde{t_i}$. The process

Algorithm 1. Distributed Algorithm for Social Welfare Maximization

```
Input: Assume the set of users in a sensing system is $\mathcal{U} = \{1, 2, \cdots, n\}$. The platform's utility function is $U(\mathbf{t})$ and the cost functions of users are $C_i(t_i), \forall i \in \mathcal{U}$.
Output: The optimal sensing time $t_i^*$ for each user and the optimal price vector $\mathbf{p} = [p_1, \cdots, p_n]$ set by the platform.
1: $l = 0$        //$l$ counts the number of iterations
2: The platform sets the same initial price $p^{(0)} = \epsilon$ sent to all users, where $\epsilon$ is a little positive real value.
3: repeat   for $l = 0, 1, \cdots$
4:     For each user, the sensing time $\widetilde{t_i}^{(l+1)}$ is computed to maximize the payoff and returned to the platform.
           $$\widetilde{t_i}^{(l+1)} = \arg\max_{t_i \in [0, \tau_i]} \quad P_i(t_i, p_i^{(l)}) = p_i^{(l)} t_i - C_i(t_i). \qquad (12)$$
5:     For the platform, after receiving $\widetilde{t_i}^{(l+1)}$, the price $p_i^{(l+1)}$ is updated and sent to user $i$.
           $$p_i^{(l+1)} = (1 - \gamma) p_i^{(l)} + \gamma \left. \frac{\partial U(\mathbf{t})}{\partial t_i} \right|_{\mathbf{t} = \widetilde{\mathbf{t}}^{(l+1)}}. \qquad (13)$$
6: until   If $\left\| \widetilde{\mathbf{t}}^{(l+1)} - \widetilde{\mathbf{t}}^{(l)} \right\|_\infty \leq \xi$, where $\xi$ is a tunable little real number.
7: $\mathbf{t}^* = \widetilde{\mathbf{t}}^{(l+1)}$, $\mathbf{p}^* = \mathbf{p}^{(l+1)}$
8: return $\mathbf{t}^*$ and $\mathbf{p}^*$.
```

repeats until the difference of sensing time $\widetilde{t_i}^{(l+1)} - \widetilde{t_i}^{(l)} (\forall i \in \mathcal{U})$ between two consecutive iterations is extremely small. The prices for different users can be updated asynchronously. For example, price $p_i^{(l)}$ can be computed based on the sensing time $\widetilde{t_j}^{(l^{'})}$, where $j \neq i, l^{'} \neq l$.

5.3 Convergence Analysis

To ensure the convergence of a certain algorithm executed in an asynchronous way, the algorithm should be a *contraction mapping* [15]. We first introduce the contraction mapping and its property. Then, the convergence conditions of Algorithm 1 is analyzed with different relaxation parameter γ.

Contraction Mapping. Many iterative algorithms can be expressed as $x(l + 1) = \Theta(x(l)), l = 0, 1, \cdots$ where $x(\cdot) \in X$ and l denotes the number of iterations. Mapping Θ is called a *contraction* if

$$\|\Theta(x) - \Theta(y)\| \leq \kappa \|x - y\|, \quad \forall x, y \in X, \qquad (14)$$

where $\|\cdot\|$ is some norm, and $\kappa \in [0, 1)$ is called the *modulus* of Θ. Moreover, the mapping Θ is called a *pseudo-contraction* if there exists a fixed point $x^* \in X$(means $x^* = \Theta(x^*)$) and

$$\|\Theta(x) - x^*\| \leq \kappa \|x - x^*\|, \quad \forall x \in X. \tag{15}$$

The convergence property of contractions or pseudo-contractions is given in Theorem 1.

Theorem 1 (Geometric Convergence). *Suppose that mapping Θ is a contraction or a pseudo-contraction and the modulus of Θ is $\kappa \in [0, 1)$. Then, Θ has a unique fixed point x^* and the sequence $\{x(l), l = 0, 1, \cdots\}$ generated by $x(l + 1) = \Theta(x(l))$ satisfies*

$$\|x(l) - x^*\| \leq \kappa^l \|x(0) - x^*\|, \quad \forall l \geq 0, \tag{16}$$

for every choice of the initial $x(0) \in X$. In particular, $x(l)$ converges to x^* geometrically.

For simplicity, we use $u_i(\mathbf{t})$ to denote for $\frac{\partial U(\mathbf{t})}{\partial t_i}$ and $c_i(t_i)$ for $C_i^{'}(t_i)$. The second order partial derivatives of U is denoted by $\partial_{x_j x_i} U(\mathbf{t}) = \nabla_j u_i(\mathbf{t})$.

Convergence of Algorithm 1 with $\boldsymbol{\gamma = 1}$. We first define a notation $[t]_i^+$ to denote the projection of $t_i \in \Re$ onto the range $[0, \tau_i]$,

$$[t]_i^+ = \arg\max_{z \in [0, \tau_i]} |z - t_i|.$$

Briefly, the solution to (8) is equivalent to

$$\widetilde{t_i}(p_i) = [\arg\max_{t_i} p_i t_i - C_i(t_i)]_i^+ = [c_i^{-1}(p_i)]_i^+.$$

Therefore, when $\gamma = 1$, the price updating rule turns to that $p_i = u_i(\widetilde{\mathbf{t}})$. Substituting it into $\widetilde{t_i}(p_i)$ obtains

$$\widetilde{t_i} = \Theta_i(\widetilde{\mathbf{t}}) = [c_i^{-1}(u_i(\widetilde{\mathbf{t}}))]_i^+. \tag{17}$$

Algorithm 1 applied with $\gamma = 1$ eventually achieves convergence under the certain condition which is given in the following proposition.

Proposition 1. *Supposing $\gamma = 1$, if we have*

$$\sum_{j=1}^{n} \left|\partial_{t_j t_i} U(\mathbf{t})\right| < \min_{t_i} \left|c_i^{'}(t_i)\right|, \forall \mathbf{t} \in \prod_i [0, \tau_i]. \tag{18}$$

Θ_i given by (17) is a contraction. According to Theorem 1, $\{\widetilde{t_i}(p_i^{(l)})\}$ generated by Algorithm 1 converges geometrically to the optimal solution t_i^* of (7), given any initial price $p_i^{(0)}$.

Proof. For each i, a function g_i is defined as follows,

$$g_i(r) = c_i^{-1}(u_i(z(r))) = c_i^{-1}(u_i(rx + (1 - r)y)), \quad r \in [0, 1],$$

where $g_i(r)$ is differentiable. We have

$$\begin{aligned}|\Theta_i(x)-\Theta_i(y)| &= \left|[c_i^{-1}(u_i(x))]_i^+ - [c_i^{-1}(u_i(y))]_i^+\right| \\ &\le |g_i(1)-g_i(0)| = \left|\int_0^1 \frac{\mathrm{dg_i(r)}}{\mathrm{dr}}\right| \\ &\le \int_0^1 \left|\frac{\mathrm{dg_i(r)}}{\mathrm{dr}}\right| \le \max_{r\in[0,1]} \left|\frac{\mathrm{dg_i(r)}}{\mathrm{dr}}\right|,\end{aligned}$$

where the first inequality is because $\left|[x_i]_i^+ - [y_i]_i^+\right| \le |x_i - y_i|$ for all $x_i - y_i \in \Re$. Furthermore, applying the chain rule, we have

$$\begin{aligned}\left|\frac{\mathrm{dg_i(r)}}{\mathrm{dr}}\right| &= \left|\sum\nolimits_{j=1}^n \nabla_j c_i^{-1}\left(u_i(rx+(1-r)y)\cdot(x_j-y_j)\right)\right| \\ &\le \left|(c_i^{-1})'(u_i(z(r)))\right| \cdot \sum\nolimits_{j=1}^n |\nabla_j u_i(z(r))| \cdot |x_j - y_j|\,.\end{aligned}$$

If condition (18) holds, we have

$$\sum_{j=1}^n |\nabla_j u_i(x)| < \min_{x_i} \left|c_i^{'}(x_i)\right| \le \left|c_i^{'}(c_i^{-1}(u_i(x)))\right| = \frac{1}{\left|(u_i^{-1})'(c_i x)\right|},$$

for all $x \in \prod_i [0,\tau_i]$. Therefore, there exists a number $\kappa < 1$ which enables that

$$\left|\frac{\mathrm{dg_i(r)}}{\mathrm{dr}}\right| \le \kappa \max_j |x_j - y_j| = \kappa \|x-y\|_\infty, \quad \forall r \in [0,1].$$

Therefore, we have that

$$\|\Theta_i(x)-\Theta_i(y)\|_\infty \le \kappa\|x-y\|_\infty, \quad \forall x,y \in \prod_i [0,\tau_i],$$

which shows Θ_i is a contraction with modulus κ respect to the maximum norm. Therefore, the proposition is proved.

Convergence of Algorithm 1 with $\gamma < 1$. We relax the value of γ by considering the convergence condition with $\gamma < 1$. In this situation, we suppose that $\widetilde{\mathbf{t}}(\mathbf{p}^{(l)})$ generated in each iteration l is always within the range $\prod_i [0,\tau_i]$. Therefore, we can rewrite the solution to (8) as $\widetilde{t_i}(p_i) = c_i^{-1}(p_i)$, which has a little difference from the $\gamma = 1$ situation. Combining it with $p_i = (1-\gamma)p_i + \gamma u_i(\widetilde{\mathbf{t}})$, we obtains

$$\widetilde{t_i} = \Theta_i(\widetilde{\mathbf{t}}) = c_i^{-1}((1-\gamma)c_i(\widetilde{t_i}) + \gamma u_i(\widetilde{\mathbf{t}})). \tag{19}$$

Applied with (19), Algorithm 1 can achieve its optimum under the convergence condition given in the following proposition.

Proposition 2. *Set $\widetilde{t_i}(p_i^{(0)})$ to be 0 for all i or τ_i for all i. If we have*

$$\begin{cases} 0 \leq \gamma \leq \left(1 + \frac{\partial_{t_j t_i} U(\mathbf{t})}{u_i^{'}(t_i)}\right)^{-1} \\ \sum_{j \neq i} \partial_{t_j t_i} U(\mathbf{t}) \leq 0 \end{cases}, \forall \mathbf{t} \in \prod_i [0, \tau_i]. \tag{20}$$

Θ_i given by (19) is a pseudo-condition for all t_i between $t_i(p_i(0))$ and t_i^*. According to Theorem 1, $\{\widetilde{t_i}(p_i^{(l)})\}$ generated by Algorithm 1 converges geometrically to the optimal solution t_i^* of (7).

Proof. For each i, a function g_i is defined as follows,

$$g_i(r) = c_i^{-1}((1-\gamma)c_i(z(r))) + \gamma u_i(z(r)),$$

where $z(r) = rx + (1-r)y$.

Supposing $x < x^*$, we will show that $x_i < \Theta_i(x) \leq x_i^*$. We have

$$\Theta_i(x) - \Theta_i(x^*) = g_i(1) - g_i(0) = \int_0^1 \frac{\mathrm{dg_i(r)}}{\mathrm{dr}} \mathrm{dr},$$

where $\frac{\mathrm{dg_i(r)}}{\mathrm{dr}}$ is given by

$$\begin{aligned} \frac{\mathrm{dg_i(r)}}{\mathrm{dr}} &= (\mathrm{c}_i^{-1})^{'} ((1-\gamma)c_i(z(r)) + \gamma u_i(z(r))) \cdot \left((1-\gamma)c_i^{'}(z(r)) \cdot (x_i - x_i^*) + \gamma \sum_j \nabla_j u_i(z(r)) \cdot (x_j - x_j^*)\right) \\ &= \left(c_i^{'}(\Theta_i(z(r)))\right)^{-1} \cdot \left(\left((1-\gamma)c_i^{'}(z(r)) + \gamma \nabla_i u_i(z(r))\right) \cdot (x_i - x_i^*) + \gamma \sum_{j \neq i} \nabla_j u_i(z(r)) \cdot (x_j - x_j^*)\right). \end{aligned}$$

Because C_i is strictly convex, $c_i^{'}(x) > 0$. Therefore, under condition (20), if $x < x^*$, we have $\Theta_i(x) \leq x_i^*$, from which we obtain

$$c_i(\Theta_i(x)) = (1-\gamma)c_i(x_i) + \gamma c_i(x) \leq c_i(x_i^*) = u_i(x^*) < u_i(x),$$

which leads to $c_i(x_i) < u_i(x)$ and thus

$$c_i(x_i) < (1-\gamma)c_i(x_i) + \gamma u_i(x), \quad \forall \gamma > 0.$$

Applying $c_i^{-1}(\cdot)$ to both sides yields $x_i < \Theta_i(x)$. Therefore, there exists a number $\kappa < 1$ enables that $|\Theta_i(x) - x_i^*| \leq \kappa\, |x_i - x_i^*|$, which is equivalent to

$$\|\Theta_i(x) - x^*\|_\infty \leq \kappa \|x - x^*\|_\infty, \quad \forall x_i \in [0, x_i^*].$$

In other words, Θ_i is a pseudo-contraction in $[0, x_i^*]$.

Similarly, if $x > x^*$, we can get Θ_i is a pseudo-contraction in $[x_i^*, \tau_i^*]$. Therefore, the proposition is proved.

6 Performance Evaluation

In this section, we perform simulations to evaluate the performance on maximizing social welfare of the proposed algorithm.

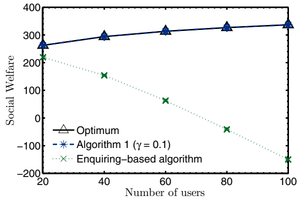


Fig. 2. Social welfare comparison of different algorithms with varying the number of users

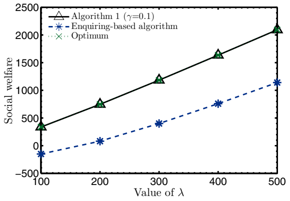


Fig. 3. Social welfare comparison of different algorithms with varying λ

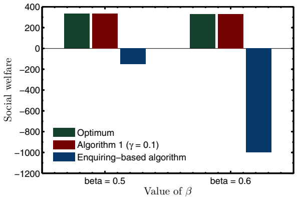


Fig. 4. Social welfare comparison of different algorithms with varying β

6.1 Methodology and Setup

We compare the proposed algorithm with a baseline algorithm, which is named *enquiring-based pricing algorithm.*

Enquiring-Based Pricing Algorithm. In this algorithm, the platform provides the same price for all users to keep fairness. As the same as Algorithm 1, the cost functions of users are unknown to the platform. The platform first notifies an initial price p_0 to all the users to enquire the promised sensing time. Next, each user returns a sensing time $\widetilde{t_i}(p_0)$ by maximizing the payoff according to (8). The platform simply assumes that the relation between the price and the sensing time is linear, i.e., $t_i = a_i \cdot p$. Therefore, a_i is obtained by the platform based on p_0 and $\widetilde{t_i}(p_0)$. Then, the platform makes the decision on the final price $\widehat{p}$ via maximizing the profit given by (5). After receiving the final price $\widehat{p}$, each user will compute the real sensing time $\widehat{t_i}$ based on (8).

We also compare the social welfare computed by Algorithm 1 with the optimal value. The optimal value is obtained by directly solving the optimal problem in (7) with the knowledge of social welfare function. For simplicity, the values of β_i for different users are the same, denoted as β. In our simulations, we vary three parameters: the number of sensing users n, α and β in our models, which influence the results of the social welfare. The default values of other parameters are set as follows. The initial prices of both algorithms are set to 1.0. α_i and W_i are normally distributed random numbers within $(0.5, 1.0)$ and $(0, 1)$, respectively. $\beta = 0.5$, $\lambda = 100$, and $\xi = 0.001$.

6.2 Comparison of Social Welfare

We first evaluate the performance on the social welfare of two algorithms under different size of users. As shown in Fig. 2, the social welfare produced by Algorithm 1 increases with more users, which is the same as the optimal value all the time. However, the social welfare produced by the enquiring-based algorithm becomes worse and worse as the size of users is larger, which even falls below zero at last. Algorithm 1 has an average social welfare 90.0% higher than that of the enquiring-based algorithm.

We next evaluate the performance on the social welfare of two algorithms under different system parameter λ, which is positive related with the utility of the platform. Fig. 3 shows that the social welfare computed by Algorithm 1 keeps the same with the optimum as λ varies. When $\lambda = 500$, the average social welfare produced by Algorithm 1 is 83.5% higher than that of the the enquiring-based algorithm.

Finally, the performance on the social welfare of two algorithms under different system parameter β, which has positive relation with the cost of users. Fig. 4 shows that the social welfare computed by Algorithm 1 keeps the same with the optimum as β varies. With different β, the social welfare computed by Algorithm 1 is stable, while that computed by the enquiring-based algorithm varies large.

7 Conclusion

This paper has focused one the problem of maximizing the system-wide social welfare in a participatory smartphone sensing system. In such a system, the cost of each user is private information, which prevents the application of a centralized algorithm to be executed by the platform. In this paper, we have proposed *the novel price-based decomposition framework*, which provides an effective incentive mechanism for stimulating smartphone users to contribute to the participatory smartphone sensing. With the framework, we have converted the original difficult problem to an optimal pricing problem. We have designed the distributed *step-size-free* price-updating algorithm which can successfully approach to the system-wide optimum. Extensive simulations have been performed and the results have confirmed that that our framework and the distributed algorithm can achieve the optimal social welfare of the participatory smartphone system.

References

1. Burke, J., Estrin, D., Hansen, M., et al.: Participatory sensing (2006)
2. Lane, N.D., Miluzzo, E., Lu, H., et al.: A survey of mobile phone sensing. IEEE Communications Magazine 48(9), 140–150 (2010)
3. Chatzimilioudis, G., Konstantinidis, A., Laoudias, C., et al.: Crowdsourcing with smartphones (2012)
4. Mun, M., Reddy, S., Shilton, K., et al.: Peir, the personal environmental impact report, as a platform for participatory sensing systems research. In: Proc. ACM MobiSys, pp. 55–68 (2009)
5. Stevens, M., D'Hondt, E.: Crowdsourcing of pollution data using smartphones. In: Proc. Workshop on Ubiquitous Crowdsourcing (2010)
6. Kanhere, S.S.: Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces. In: Hota, C., Srimani, P.K. (eds.) ICDCIT 2013. LNCS, vol. 7753, pp. 19–26. Springer, Heidelberg (2013)
7. Yan, T., Marzilli, M., Holmes, R., et al.: mcrowd: a platform for mobile crowdsourcing. In: Proc. ACM SenSys, pp. 347–348 (2009)
8. Sheng, X., Xiao, X., Tang, J., Xue, G.: Sensing as a service: a cloud computing system for mobile phone sensing. In: Proc. IEEE Sensors, pp. 1–4 (2012)

9. Boulos, M.N.K., Resch, B., Crowley, D.N., et al.: Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, ogc standards and application examples. Health Geographics 10(1), 67 (2011)
10. Lee, J.S., Hoh, B.: Sell your experiences: a market mechanism based incentive for participatory sensing. In: Proc. IEEE PerCom, pp. 60–68 (2010)
11. Lee, J.S., Hoh, B.: Dynamic pricing incentive for participatory sensing. Pervasive and Mobile Computing 6(6), 693–708 (2010)
12. Jaimes, L.G., Vergara-Laurens, I., Labrador, M.A.: A location-based incentive mechanism for participatory sensing systems with budget constraints. In: Proc. IEEE PerCom, pp. 103–108 (2012)
13. Yang, D., Xue, G.,, X.F., et al.: Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing. In: Proc. ACM MobiCom, pp. 173–184 (2012)
14. Koutsopoulos, I.: Optimal incentive-driven design of participatory sensing systems. In: Proc. IEEE Infocom (2013)
15. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and distributed computation. Prentice Hall (1989)

An Optimal Solution for Round Rotation Time Setting in LEACH

Hongyan Zhang, Xin Li, and Xiumei Fan

School of Computer Science,
Beijing Institute of Technology, Beijing, China
{2120111219,xinli,xmfan}@bit.edu.cn

Abstract. There have been many protocols proposed for wireless sensor networks (WSN) where the energy awareness is an essential issue. Low Energy Adaptive Clustering Hierarchy (LEACH) is a widely adopted cluster-based structure for the energy-aware WSN, which utilized a Time Division Multiple Access(TDMA)-based MAC protocol to maintain balanced energy consumption and has shown effectiveness in prolonging the lifetime of sensors. However the related parameters setting in LEACH is the tricky and essential part for achieving good performance e.g., the number of clusters, the rotation time for each round. In literature, researchers used the empirical value as the round rotation time to obtain good performance. In this paper, we use Voronoi region to describe the distribution form of the cluster head and its members to conduct the theoretically optimal solution of the duration for each round. The experimental results show that using our suggested setting for the round rotation time is much more effective and efficient than the conventional LEACH with the empirical settings in terms of energy saving and the network surviving, and the amount of data delivered to the base station.

Keywords: Energy Dissipation, Cluster, Voronoi Tessellation, Round Rotation Time.

1 Introduction

Wireless Sensor Network (WSN) consists of hundreds or thousands of sensors which can communicate with each other to detect and process data cooperatively and gather information from the environment in which they are deployed. WSN now has been widely adopted in many applications such as military defense, environment monitoring, industrial sense, control applications and smart building etc.[1]. Despite its well-known success, WSN still suffer from its non-lasting lifetime as the sensors are usually small-sized sensing devices with limited power provided by the battery. Thus, how to efficiently utilize the limited power and extend the network lifetime is an important issue. In literature, a bunch of energy-efficient protocols have been proposed to prolong WSN's lifetime. One of the promising directions is the cluster-based mechanism by which the sensors can group themselves into clusters to achieve the balanced energy consumption to prolong the network's lifetime.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 366–376, 2013.

Low Energy Adaptive Clustering Hierarchy (LEACH) proposed by Heinzelman et al.[2] is one of the most popular cluster-based energy-efficient communication protocol for wireless sensor networks. It reduces the global energy consumption by clustering the nodes and distributing the load to all the clusters. The operation of LEACH is divided into rounds and each round consists of a set-up phase and a steady phase. The set-up phase consists of selecting cluster-head, forming clusters, creating TDMA schedule. In the steady phase, the cluster heads receive message from their members and send the aggregated message to the base station. However, many parameters in LEACH can affect the network performance greatly which need to be optimized, e.g., the threshold, the number of clusters. Accordingly, there are some researchers working in polishing LEACH to improve its performance.

The optimized threshold are proposed in[3,4,5] to select the cluster heads in an efficient way. Also, the optimal solution for the proper number of clusters are investigated in[6]. The authors proposed to calculate the range of the optimum number of clusters by property of that the optimum number of clusters is inversely proportional to the distance between the Cluster Heads (CH) and the base station. While EDGA[7] is based on the weighted election probabilities of each node to choose cluster heads. EDGA adopted intra-cluster coverage to solve the area coverage problem in a cluster range. With the weighted election probabilities, EDGA prolonged the network lifetime. An other optimized version of LEACH, LEACH-C is proposed in[8]. During the LEACH-C's setup stage, the base station receives the location and energy of each node in the network. According to these information, the base station selects the best cluster heads. The simulation in LEACH-C delivers 40% more data per unit energy than the conventional LEACH. In[9] the authors extend LEACH's stochastic cluster heads selection algorithm by adding a deterministic component and change the probability of a node being a cluster head by the remaining energy of sensor nodes.

Although a lot of work focus on optimizing LEACH protocol, round rotation time is rarely discussed in literature. The rotation time setting is critical. If each round lasts too long, the cluster heads stay active all time, the energy of the cluster head may drain quickly. If the round rotation time is too short, the network rebuild the clusters frequently and then most of the energy will be wasted in the set-up phase instead of transmitting the data in the steady phase. In this paper, we focus on how long each round should last before the network restart another round. We adopt Voronoi region to describe the distribution form of the cluster heads and their cluster members to obtain the optimal round rotation time. The simulation results show that our conducted best round rotation time for LEACH is much more energy-efficient, which successfully prolongs the network lifetime and delivers more data to the base station comparing with the conventional LEACH.

The rest of the paper is organized as follows. Section 2 presents a brief description of LEACH protocol, Section 3 provides the details of our proposed method to compute the best round rotation time. The experimental results based on several benchmark problems are reported in Section 4. The conclusion and possible extensions are presented in Section 5.

2 Background

LEACH is a low-power adaptive clustering topology protocol. To ensure energy consumption balanced, LEACH randomly and periodically select the nodes to be the cluster heads among all the nodes. The operation of LEACH is divided into rounds and each round consists of a set-up phase and a steady phase. During the setup phase, nodes communicate with each other and are organized into clusters with accomplishment of the cluster heads selection. More detailedly, LEACH generates a random number of 0 to 1 for each node, if the random number is less than the threshold value $T(n)$, the node will be chosen as the cluster head of the current round. $T(n)$ is defined as follows:

$$T(i) = \begin{cases} \frac{p}{1-p*(rmod1/p)} & if\ i \in G \\ 0 & otherwise \end{cases} \tag{1}$$

Where p is the percentage of the expected cluster heads in the total sensor nodes and r is the number of rounds that have already passed. And G is the set of the sensor nodes which have not been selected as cluster heads yet in the last 1/p rounds. Eq.(1) ensures that each node is selected as a cluster head only once during 1/p rounds, therefore all nodes have the equal chance to be a cluster head and thus energy could be consumed fairly. After selected as the cluster head, the nodes broadcast the advertisement message to the other nodes. The other nodes will join the cluster based on the its received signal's strength. If the strength of advertisement signal sent from one cluster head is much stronger than the others, the node would volunteered to be the member of this cluster head which implies that the senor nodes tend to join the cluster which cluster head is nearest to them. In steady phase, each cluster head sets up Time Division Multiple Access(TDMA) schedules for all its cluster members, then every non-cluster-head nodes in this cluster will send the data to its cluster head according to the predefined TDMA schedule. Furthermore, the radio of each non-cluster-head node is turned off until its transmission time comes. Before sending data to the base station, the cluster head aggregate all the data. Via the random selection of the cluster heads, LEACH managed to evenly distribute the massive energy consumption to all the nodes in the network. And the consumption of the energy follows a uniform distribution.

3 An Optimal Round Rotation Time Conduction for LEACH

In this section, we discuss how to conduct the optimal round rotation time setting for LEACH. As mentioned before, the time setting for the round rotation is essential to LEACH. If one round lasts too long, the cluster head will be died soon. On the other hand, if one round lasts too short, the topology of the clusters will change frequently and a lot of energy is wasted in set-up phase instead of delivering the data. Due to the limited energy, the duration of the steady phase should be much longer than that of the set-up phase to accomplish the data transmission. And the set-up phase will cost a trivial energy dissipation.

Assume there are N nodes uniformly distributed in an $M \times M$ field. And the number of clusters is k. There are $\frac{N}{k}$ nodes in each cluster on average and $\frac{N}{k}-1$ nodes as none-cluster-head. Due to the base station is usually far away from the field, we consider the distance to the base station is greater than the cross-over distance and then energy dissipation follows the multi-path model[7]. The energy of each cluster head is used to receive the data from its cluster members, aggregate the received data and send the aggregated data to base station. Thus the energy dissipated in the cluster head nodes $i(i = 1, \cdots N)$during a single frame is as below:

$$\begin{aligned} E_{ch} &= lE_{elec}(\tfrac{N}{k}-1) + lE_{DA}\tfrac{N}{k} + lE_{elec} + l\varepsilon_{amp}d_i^4 \\ E_{ch} &= lE_{elec}\tfrac{N}{k} + lE_{DA}\tfrac{N}{k} + l\varepsilon_{amp}d_i^4 \end{aligned} \tag{2}$$

Where l is the number of bits in each message, E_{elec} is the energy dissipated per bit, E_{DA} is the energy for aggregating the data, d_i represents the distance between the cluster head node i and the base station. The non-cluster-head nodes are only for transmitting data to its corresponding cluster head. In the common scenario, we consider the distance from non-cluster-head nodes to their corresponding cluster heads is relatively small, then the energy dissipation follows the Friss free-space model[7]. So the energy consumption of each non-cluster-head is defined as below:

$$\begin{aligned} E_{non-CH} &= lE_{elec} + l\varepsilon_{fs}d_{toCH}^2 \\ E(E_{noch}) &= lE_{elec} + l\varepsilon_{fs}E(d_{toCH}^2) \end{aligned} \tag{3}$$

where d_{toCH} is the distance from the non-cluster nodes to their corresponding cluster heads. To evaluate Eq.(3), we should know the value of d_{toCH}. Recall that, after being selected to be the cluster heads, nodes broadcast an advertisement message to the rest of the nodes. The rest of nodes join the cluster based on the received advertisement's signal strength. The signal strength is determined by the distance between the nodes. The closer the distance, the stronger the signal is. Then the distance from the non-cluster-nodes and corresponding cluster head is much closer than the other cluster heads. And this is basically the idea of how the Voronoi tessellation refine its region.

Given an open set $\Omega \in R^N$,the set $\{V_i\}_{i=1}^k$ is called a tessellation of Ω (The Voronoi tessellation [10]) if $V_i \subset \overline{\Omega}$ for i=1,...k,$V_i \cap V_j = \Phi$ for $i \neq j$ and$\cup_{i=1}^k \overline{V_i} = \overline{\Omega}$. We use $d(x, y)$ represent the distance between x and y. Given points $\{z_i\}_{i=1}^k$ belonging to the closed set $\Omega \in R^N$, the Voronoi region $\hat{V_i}$ corresponding to the point z_i is defined as:

$$\hat{V_i} = \{x \in \Omega \mid d(x, z_i) < d(x, z_j) \qquad for \quad j = 1, ...k, j \neq i\}$$

The set $\{\hat{V}_i\}_{i=1}^{k}$ is called a Voronoi tessellation of Ω. Suppose the position of each cluster head (x^*, y^*) is the center of its corresponding Voronoi region $\hat{V}_i$ and any non-cluster node position is (x, y). Since the nodes are uniformly distributed in the area of clusters, the covered area of each cluster is $\frac{M^2}{k}$. As shown in Figure 1, we consider each cluster corresponds to a Voronoi tessellation. Then each Voronoi tessellation is a $\sqrt{\frac{M^2}{k}} \times \sqrt{\frac{M^2}{k}}$ square. Accordingly we obtain the density for the node coordination is $\rho[x] = \rho[x^*] = \rho[y] = \rho[y^*] = \sqrt{\frac{k}{M^2}}$. Then we have $E[x] = E[x^*] = E[y] = E[y^*] = \frac{1}{2}\frac{M^2}{k}$ and $E[x^2] = E[{x^*}^2] = E[y^2] = E[{y^*}^2] = \frac{1}{3}\frac{M^2}{k}$. The expected distance between a cluster head and its member is then achieved as below

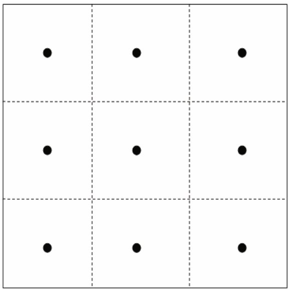


Fig. 1. Voronoi tessellation separation

$$\begin{aligned} E[d_{toCH}^2] &= E[(x - x^*)^2 + (y - y^*)^2] \\ &= E(x^2 + {x^*}^2 - 2xx^* + v^2 - 2yy^*) \\ &= 4E(x^2) - 4E(x)E(x^*) \\ &= \frac{M^2}{3k} \end{aligned} \quad (4)$$

With substituting Eq.(4) into Eq.(3), we obtain:

$$E_{non-CH} = lE_{elec} + l\varepsilon_{fs}\frac{M^2}{3k} \quad (5)$$

The total energy consumed by each node per round depends on the average number of frames per round can be written as:

$$E_{CH|round} = N_{frames|round} \times E_{CH|frame} \quad (6)$$

$$E_{non-CH|round} = N_{frames|round} \times E_{non-CH|frame} \quad (7)$$

Where $E_{CH|round}$ is the energy which cluster heads consumed to receive data from their cluster members, aggregate data and communicate with the base station, $N_{frmaes|round}$ is the average number of frames per round. $E_{non-CH|round}$ is the energy used by non-cluster-head nodes to send data to their corresponding cluster heads. The minimum time for each round should be sufficient enough for allowing a node to be cluster-head at least once and non-cluster-head many times, before the energy is used up. For example, if there are $\frac{N}{k}$ rounds in the network operation, we should guarantee that the energy of each node is able to sustain playing the role of cluster head once and cluster member $\frac{N}{k} - 1$ times. According to Eq.(2)(4)(6)(7), we obtain:

$$\begin{aligned} E_{total} &= E_{CH|round} + \left(\frac{N}{k} - 1\right) E_{non-CH|round} \\ &= N_{frames|round}\left(lE_{elec}\frac{N}{k} + lE_{DA}\frac{N}{k} + l\varepsilon_{amp}d_i^4\right) \\ &+ N_{frames|round}\left(\frac{N}{k} - 1\right)\left(lE_{elec} + l\varepsilon_{fs}\frac{M^3}{3k}\right) \end{aligned} \tag{8}$$

$$\begin{aligned} &N_{frames|round} \\ &= \frac{E_{start}}{l[(E_{elec}\frac{N}{k} + E_{DA}\frac{N}{k} + \varepsilon_{amp}d_i^4)((\frac{N}{k} - 1)(E_{elec} + \varepsilon_{fs}\frac{M^2}{3k}))]} \end{aligned} \tag{9}$$

In WSN, the bitrate is a predefined as a constant value R_b , the time for transmitting one bit is then denoted as t_{data},$t_{data} = \frac{l}{R_b}$. As there are $\frac{N}{k}$ nodes in the cluster, and the total time for all nodes in the cluster to transmit one message is:

$$t_{frame} = \frac{N}{k}\frac{l}{R_b}$$

The time of each round is then

$$\begin{aligned} t_{round} &= N_{frames|round} \times t_{frame} \\ &= \frac{Nl}{kR_b}\frac{E_{start}}{[(E_{elec}\frac{N}{k} + E_{DA}\frac{N}{k} + \varepsilon_{amp}d_i^4) + (\frac{N}{k} - 1)(E_{elec} + \varepsilon_{fs}\frac{M^2}{3k})]} \end{aligned} \tag{10}$$

To make the network more energy-efficient, we should consider all the clusters and prolong the time for one round as long as possible. Thus the optimal round rotate time should be:

$$\begin{aligned} &t_{rotation} \\ &= max(\frac{Nl}{kR_b}\frac{E_{start}}{[(E_{elec}\frac{N}{k} + E_{DA}\frac{N}{k} + \varepsilon_{amp}d_i^4) + (\frac{N}{k} - 1)(E_{elec} + \varepsilon_{fs}\frac{M^2}{3k})]}) \end{aligned} \tag{11}$$

4 Simulation and Performance Analysis

For the experiments, we use NS2[11] to simulate a WSN with 50, 100, 200 nodes respectively. We observed the network lifetime for various choices of round rotation time in the network and show the round rotation time we derived successfully prolong the network lifetime best. Also, we plotted the changes of energy consumption and the number of alive nodes over time specifically for a 100 nodes network with different round rotation time settings. The experimental results illustrated the round rotation time we conducted is much more energy-efficient and keep more nodes alive.

The parameters in our experiments are showed in Table 1:

Table 1. The Corresponding Parameters Settings in our Simulation

Description	Parameter	Value
Radio electronics energy	E_{elec}	$50nJ/bit$
Compute energy for data aggregation	E_{DA}	$5nJ/bit$
Base station location	L_{BS}	(50,175)
Node number	N	50 100 200
Network size	$M \times M$	100×100
The proportion of cluster head	$\frac{k}{N}$	5%
Data size	l	$4000bits$
Bitrate	R_b	$1Mbps$
Initial energy	E_{star}	2mJ
Radio amplifier energy	ε_{amp}	$0.0013pJ/bit/m^4$
	ε_{fs}	$10pJ/bit/m^2$

According to Eq.(11), we can see that the optimal round rotation time is related to the proportion of cluster head ($\frac{k}{N}$), no matter what the nodes number is(Seen from the parameter setting in Table, the last term in Eq.(11) is much smaller than the former ones). With the settings in table 1, we obtain the optimal rotation time $t_{rotation} = 20secs$ in all three different networks.

In WSN there are different interpretation for the network's death. Here, we adopted a popular one to consider the network is down when the number of alive nodes is less than the number of clusters. Figure 2 shows the networks lifetime reaches the peak when $t_{rotation} = 20$ in three different cases, which is consistent with our conducted optimal value. We can also see that the lifetime of network with 100 nodes is much longer than other two cases. In most cases, the lifetime of network with 200 nodes is longer than 50-nodes network. That is because that the density of nodes is also another factor to affect the lifetime of network for LEACH which however will not be discussed further in this paper.

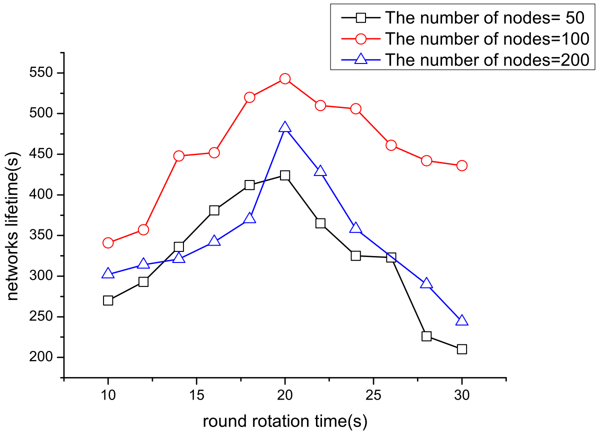


Fig. 2. Network Lifetime in Three different Networks

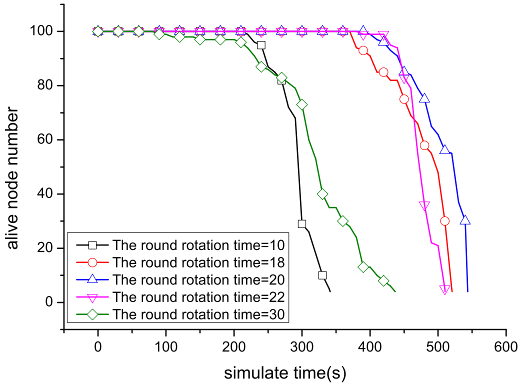


Fig. 3. The Number of Alive Nodes in a Network with 100 Sensors

While Figure 3 shows the number of alive nodes with different round rotation time setting in the 100-nodes network. We can see that, with the round rotation time setting as 20 seconds, the alive number is reduced more slowly over time than that of others. And obviously the first dead node comes up later than the others. When we set the round rotation time too long, say 30 seconds, we can see that the first dead node comes up much earlier and the lifetime is shorter that others. This is because in steady phase cluster heads keep alive all the time and energy get dissipated soon. If we set the round rotation time too short, say 10 seconds, the energy get drained also soon. This is because the network changes cluster heads too frequently and then the energy dissipated in the setup phase is not trivial any more as the number of rounds increases a lots.

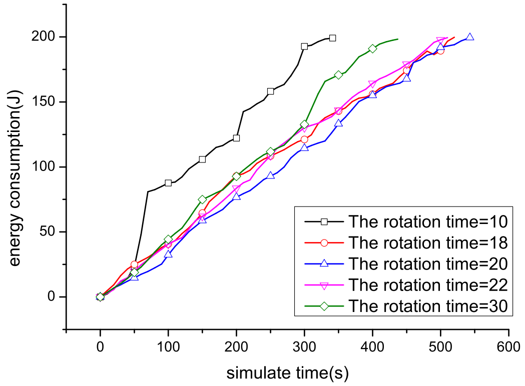


Fig. 4. Energy Consumption in 100-nodes Network

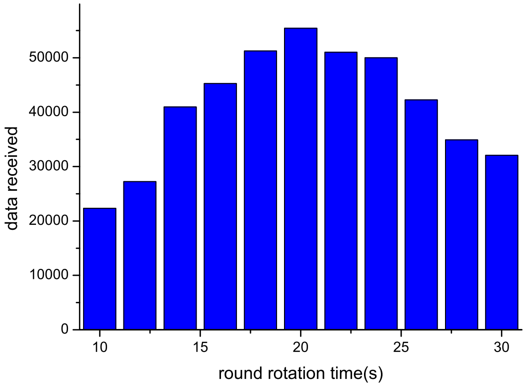


Fig. 5. data amount received at base station

We also compared the energy consumption with different round rotation time in the 100-nodes network. From Figure 4, we can see that the round rotation time we obtained (the rotation time =20) consumes lest energy at most of the time. Again, on one hand we can see when the round rotation time is too long the energy reduces quickly. On the other hand, if it is too short, the energy reduces most quickly, as the proportion of set-up phase in one round increased and now set-up phase need more energy than steady phase per second.

From Figure 5, we observe that the data amount received at base station is the most when the round rotation time is 20 seconds. If the round rotation time is too long, the nodes drain its energy quickly, then the data generates at cluster reduce which lead to the data amount received at base station reduce. If the round rotation time is too short, the time for steady phase reduces and more time is wasted in set-up phase which is useless for data transmitting. So the data amount received at base station sharply reduces.

5 Conclusion and Future Work

This paper presents an efficient way to obtain the best round rotation time in LEACH protocol. We adopts the Voronoi region to describe the distribution form of cluster head node and its members and to estimate the area coverage for a cluster. Instead of trying different time settings to obtain an empirical "good" rotation time setting, with our proposed method, we could directly draw the proper rotation time for LEACH in any networks with different parameter settings. The experimental results show the effectiveness and efficiency of our proposed method. With our conducted rotation time setting, the network performs best in terms of saving energy consumption, prolonging the network's life time as well as increasing the amount of successfully delivered data to the base station. Note that in this paper we only consider the sensor nodes are physically clustered in LEACH. When the network is logically partitioned into clusters, Voronoi region can not be directly adopted to estimate the distributions of the clusters which we found it interesting and worth for further investigation in the future work. Currently, what we derived is a "static" round rotation time according to the initial energy of the nodes which will stay same during the whole process. While energy dissipates round after round, a potential improvement is to dynamically compute the proper rotation time according to the current remained energy of nodes round after round, which will be further investigated in the future work.

Acknowledgments. This work has been partially supported by the Ph.D. Programs Foundation of Ministry of Education(Grant No.20111101120030), Beijing Institute of Technology Foundation(Grant No.20110742010), Beijing Institute of Technology Research Base Foundation(Grant No. 2012CX02021), National Natural Science Foundation of China(Grant No.61272509). The authors thank Mr. Maocai Fu for his help on the experimental settings. And the authors are much obliged to the editor and the anonymous reviewers for their insightful and detailed comments.

References

1. Arampatzis, T., Lygeros, J., Manesis, S.: A survey of applications of wireless sensors and wireless sensor networksIntelligent Control. In: Proceedings of the 2005 IEEE International Symposium, pp. 719–724. IEEE Press (2005)
2. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-Efficient communication protocol for wireless microsensor networks. In: 33rd Annual Hawaii International Conference. IEEE Press (2002)
3. Handy, M.J., Haase, M., Timmermann, D.: Low energy adaptive clustering hierarchy with deterministic cluster-head selection. In: International Workshop on 4th Mobile and Wireless Communications, Stockholm, Sweden, pp. 368–372 (2002)
4. Tao, L., Qing-Xin, Z., Luqiao, Z.: An improvement for LEACH algorithm in wireless sensor network. In: The 5th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 1811–1814. IEEE (2010)

5. Gajjar, S.H., Dasgupta, K.S., Pradhan, S.N.: Lifetime improvement of LEACH protocol for Wireless Sensor Network. In: Nirma University International Conference on Engineering (NUiCONE), pp. 1–6. IEEE (2012)
6. Heimlman, W., Chandrakasan, A., Balakrishnan, H.: An application-Specific protocol architecture for wireless microsensor networks. In: IEEE Transaoction on Wireless Neworking, pp. 660–670. MIT, Cambridge (2002)
7. Mao, Y., Liu, Z., Zhang, L., Li, X.: An effective data gathering scheme in heterogeneous energy wireless sensor networks. In: International Conference on Computational Science and Engineering, CSE 2009, pp. 338–343. IEEE (2009)
8. Heimlman, W., Chandrakasan, A., Balakrishnan, H.: An application-Specific protocol architecture for wireless microsensor networks. In: IEEE Transaoction on Wireless Neworking, pp. 660–670. MIT, Cambridge (2002)
9. Handy, M.J., Haase, M., Timmermann, D.: Low energy adaptive clustering hierarchy with deterministic cluster-head selection. In: 4th International Workshop on Mobile and Wireless Communications Network, pp. 368–372. IEEE (2002)
10. Yang, H., Sikdar, B.: Optimal cluster head selection in the leach architecture. In: Performance, Computing, and Communications Conference, IPCCC 2007, pp. 92–100. IEEE (2007)
11. UCB/LBNL/VINT Network Simulator - ns (Version 2), `http://www-mash.cs.berkeley.edu/ns/`

UPC-MAC: A Power Control MAC Protocol for Underwater Sensor Networks*

Yishan Su[1], Yibo Zhu[2], Haining Mo[2], Jun-Hong Cui[2], and Zhigang Jin[1]

[1] School of Electronic and Information Engineering,
Tianjin University, Tianjin, P.R. China
suyishan_tj@hotmail.com, zgjin@tju.edu.cn

[2] Computer Science and Engineering Department,
University of Connecticut, Storrs, CT, USA
{yibo.zhu,haining.mo,jcui}@engr.uconn.edu

Abstract. After comparing the spatial reuse efficiency between RF based networks and acoustic based networks in terms of our newly defined metric, spatial reuse index, we found that the spatial reuse efficiency in acoustic based networks is significantly lower due to the relatively low spreading loss of acoustic signals, which further degrades the network throughput. To improve the system performance, we propose a slotted based Underwater Power Control MAC protocol (UPC-MAC), which leverages transmission power and long propagation delays to enhance the spatial reuse efficiency. UPC-MAC is a reservation based channel access scheme and makes use of long propagation delays to collect neighboring nodes' sending requests and channel information between these senders and receivers. With such information, UPC-MAC allows for concurrent transmissions by applying Nash Equilibrium to transmission power adjustment , which can be done on every sending node in a distributed way. Simulation results show that UPC-MAC outperforms Slotted FAMA in terms of network goodput by 15-20% and 35-60% respectively in two representative network scenarios.

1 Introduction

UWSNs (Underwater Sensor Networks) have gained tremendous attention in recent years because of their wide civilian and military applications such as monitoring, surveillance, navigation [1,2,3]. This motivates more research for a reliable and efficient network design, from the physical layer to the upper layer protocol stack and also cross layer design [4] of UWSNs . To date, acoustic communication is proofed to be the only practical method for long range communication in underwater [5]. However, it suffers from several limitations, including the limited bandwidth and the long propagation delay. For the bandwidth, current acoustic communication system is up to $40km\cdot$ kbps for the range-rate product;

* This work was supported in part by the National Natural Science Foundation of China through the Grants 61162003, 61202379 and Qinghai Science and Technology Project No. 2012-Z-902.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 377–390, 2013.

For propagation delay, the speed of sound underwater is approximately $1500m/s$, which is 2×10^5 times lower than the speed of radio. Both of these limitations highly degrade the network throughput.

Reservation based MAC protocols have been widely adopted in UWSNs, because they can effectively alleviate collisions and achieve a relatively high network throughput. MACA [6], MACAW [7] have been proposed to solve the hidden terminal problem while FAMA [8] and Slotted FAMA [9] has been proved to be effective to tackle the long propagation delay issue in underwater environments.

However, the network capacity depends on the achievable channel capacity at each individual link and the level of spatial reuse: the total number of concurrent transmissions that can be accommodated in the network. In [10], the authors show that these reservation based protocols are inefficient in spatial reuse. Because of a relatively low spreading loss of the acoustic signal in underwater, the interference range of an acoustic communication is larger than that of an RF communication with the same transmission range. The authors in [10] show that in UWSNs, an RTS/CTS handshake becomes ineffective when the distance between a communication pair is larger than 22% of the maximum transmission range; by contrast, in radio networks, an RTS/CTS handshake retains effective as long as a communication pair is closer than 56% of the maximum transmission range. To solve the interference range uncertainty issue, power control has been proposed and extensively studied in wireless networks, which serves the purpose of both power saving and better spatial resource utilization efficiency.

In this paper, we incorporate a power control scheme into the reservation based channel access. The purposes are in two folds. On one hand, the transmission power level can be reduced to achieve a better energy efficiency. On the other hand, the interference range can be reduced for better spatial reuse efficiency. Besides, we implement a distributed algorithm to allow every node to allocate its transmission power by collecting the channel information and the nodes sending requests. Specifically, based on the collected channel information, an optimization problem, whose objective is to maximize the overall network throughput while preventing nodes from using unnecessarily high power, can be formalized to manage the transmission power. Then, we can calculate the optimal sending power for each node by solving an NE equation for a given utility function with game theory.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the difference of spatial reuse in RF networks and UWSNs. In Section 3, we propose a slotted reservation based power control MAC protocol, UPC-MAC, which employs a game theory based power allocation algorithm. In Section 4, we provide simulation results and insight on the performance of UPC-MAC compared with a classic reservation based MAC protocol Slotted FAMA. Finally, our main conclusions and future works are drawn in Section 5.

2 Background

Power control schemes for wireless networks can be generally classified into two classes. In the first class, power control is used to control the network topology.

In the second class, it is applied on the packet level, with the transmission power dependent on senders and receivers. The second class can be further divided into two subclasses: energy oriented and throughput oriented. The following power control discussion is on the packet level and throughput oriented.

In wired networks, MAC algorithms are supposed to prevent simultaneous transmissions from happening, as much as possible, since such transmissions are bound to produce collisions. A requirement is that MAC protocols should nevertheless allow as many simultaneous and successful transmissions as possible in different parts of the network. This ability of wireless networks is known as spatial reuse [11].

In both RF and Underwater acoustic (UA) communications, spatial reuse is an efficient way to improve the throughput of a network. MAC schemes with power control allow for more concurrent transmissions and increase the spatial reuse efficiency.

However, these concurrent transmissions would cause more collisions. The range within which collisions happen is usually significantly larger than the range of successful packet reception. Here we define the interference range as a radius to a pair of sender and receiver such that if some nodes send packets within this radius, it can cause collision to an on-going transmission between the sender and the receiver. Interference range is dynamic because it depends on: 1. the number of potential interferant nodes and their transmission power; 2. channel environment; 3. transmission power of the sender and the decoding threshold of the receiver. Concurrent transmissions will face interference from each other. This situation is much worse in underwater environments because of both the relatively low acoustic signal spreading loss and the high decoding threshold. To explore the spatial reuse model based on [10] and Xu's work [12]. To achieve concurrent communications, one condition has to be met: the receiving signal to interference and noise ratio (SNIR) must be higher than a certain threshold, namely: $SINR > SINR_{threshold}$. In our model:

$$SINR = \frac{p_i \cdot h_i}{\sum_{j=1}^{N-1} p_j \cdot h_j + \sigma^2} \tag{1}$$

Here p_i/p_j is the transmission power of the original sender /interfering nodes, h_i/h_j is the channel gain between the original sender/interfering nodes and a receiver, N is the number of senders (including the original sender and interfering senders), σ^2 is the thermal noise. For simplification, we assume h_i is only related to distance and ignore thermal noise and therefore we get equation (2).

$$SINR = \frac{p_i \cdot d_i^{-k}}{\sum_{j=1}^{N-1} p_j \cdot d_j^{-k}} \tag{2}$$

Where k is the spreading exponent; d_i is the transmission range of the original sender; $max\ (d_j)|_{j=1}^{(N-1)}$ is the interference range of the receiver. In RF-based networks, research shows that $k_{RF} \approx 4$, for a two ray ground-reflection model [13]; while $k_{UA} \approx 1.5$ in a UA-based network with a mixed-exponent spreading model.

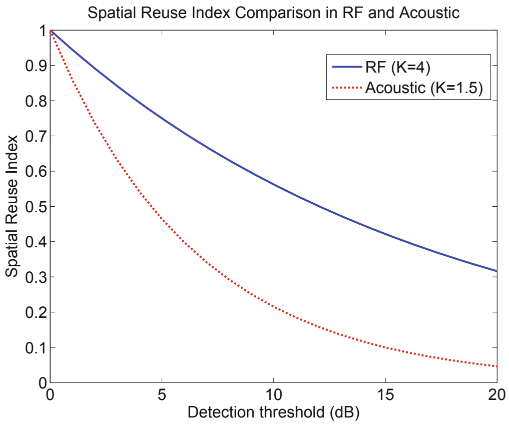


Fig. 1. Spatial Reuse Index Comparison

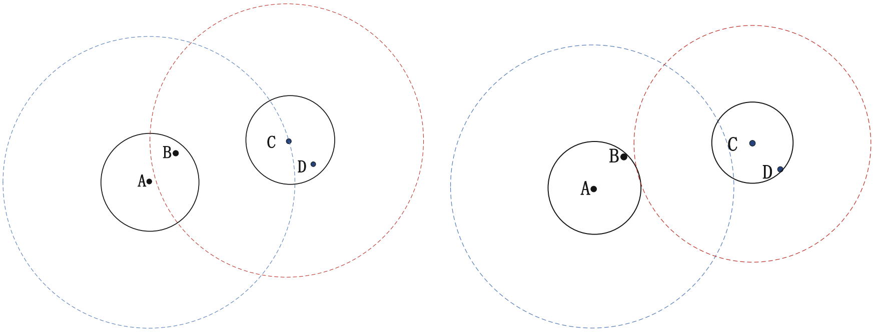
$k_{RF} > k_{UA}$ indicates that the RF signal fades faster than UA signal. Different spreading exponent reflect different spatial reuse efficiency. To quantify the spatial reuse efficiency, here we define a spatial reuse index: Ω

$$\Omega = \frac{transmission\ range}{interference\ range} \tag{3}$$

A higher Ω means a better spatial reuse efficiency. Fig. 1 shows the Ω of an RF and a UA network both involves two concurrent communication pairs. For example, if the threshold is 10dB, the collision area in UA network is 2.58 times larger than in the RF network. That is, a higher transmission power would result in a much larger interference area in a UA network higher than in an RF network. From the analysis above, we know that the spatial reuse in UWSNs is more sensitive to the transmission power.

Since acoustic signal fades lower than RF signal, the unnecessarily high transmission power can result in a larger collision area. An effective method to reduce the collision area is to avoid the unnecessarily high transmission power by implementing a power control scheme. An example is shown in Fig. 2, node A and C are senders; node B and D are the destinations of A and C respectively. The small solid circles are the transmission ranges and the larger circles are the interference range. Fig. 2(b) demonstrates the advantage of power control, which allows two concurrent transmissions without collision. For instance, by adopting an appropriate power level at node C, the transmission between node C and D will not interfere with that between node A and B.

In the following section we will introduce our new MAC protocol with throughput oriented power control algorithm.



(a) without power control

(b) with power control

Fig. 2. Two Pair Transmission

3 UPC-MAC Algorithm

In this section, we will first introduce the basic idea of UPC-MAC and then describe its workflow with an example. After that, we will formulate an optimization problem for a power control game and then derive the solution.

3.1 MAC Protocol Overview

Assumption. We assume that the channel gain is stationary during a short period, which is long enough for the transmission of a few control packets and one data packet. We also assume channel gain reciprocity, i.e. the channel gain between two terminals is the same for both directions of the transmission. This is an underlying assumption in most RTS/CTS based protocols.

UPC-MAC Overview. UPC-MAC is a reservation and slotted based MAC protocol. Here we define 4 types of packets: Request to Send (RTS), Channel State Information (CSI), DATA and Acknowledgement (ACK). Each packet (RTS, CSI, DATA, ACK) should be transmitted at the beginning of a time slot. The slot length has to be determined in a manner that avoids any data packet collision. Given by [9], the length of one slot is $\tau + \gamma + \alpha$, where τ is the maximum propagation delay; γ is the transmission time of a CSI packet and α is a guard time to compensate for possible clock drifts.

When a node A requires to send a packet, it will wait until the beginning of the next slot and then transmits an RTS packet with the maximum power if the channel is idle. This packet will be received by its destination B and all neighborhoods C and D in the transmission range within the same slot. There are two reasons why RTS is sent at the maximum power: 1. Source node has no knowledge of the position of its destination; 2. This RTS packet is needed by neighbors to collect sending request and calculate channel information. RTS packet will also include a priority which can be used for fairness, as discussed later.

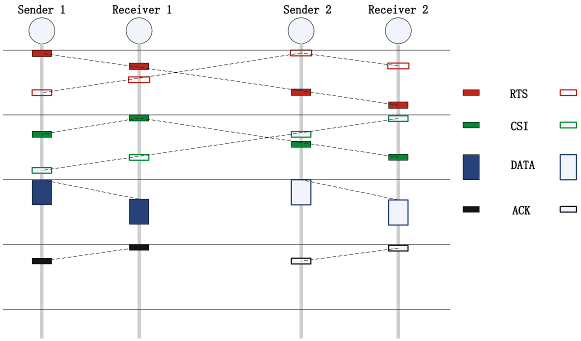


Fig. 3. UPC-MAC Overall Work Flow

All nodes that received RTS packets will calculate the channel gain, using equation (4).

$$h_i(dB) = p_r(dB) - p_s(dB) \tag{4}$$

Here, h_i is the channel gain, p_r is the receiving power of the packet, p_s is the sending power of the packet.

Then if a node is the destination of one RTS packet, it will send a CSI packet back to the sender with the maximum power level. A CSI packet contains all channel gain between a node and its neighbors. It can also be overheard by its neighboring nodes. When a source node receives all the CSI packets from its neighbors and the destination, it will collect the channel information between these nodes and itself. Then it can use the power allocation algorithm to adapt its transmission power for DATA packet if it is allowed to send the DATA packet. Otherwise, if the sender is not allowed for sending packets in the next slot, it will keep silent and wait a random back off slot to retransmit an RTS packet. When the destination receives the DATA packet, it will send an ACK packet with the same transmission power as the DATA packet. The overall workflow of UPC-MAC is illustrated in Fig. 3.

Here we use an example to show how UPC-MAC works. As shown in Fig. 4, there are six nodes in this network. At the beginning of a slot, A, B, and C wish to send a packet destined to D, E, and F, respectively. These six nodes are within the maximum transmission range of each other. Here we use $(TYPE)_{sender \to receiver}$ to denote a packet and $h_{sender \to receiver}$ to denote the channel gain between the sender and the receiver. At the beginning of the first slot, A, B, C send $RTS_{A \to D}$, $RTS_{B \to E}$, $RTS_{C \to F}$ with the maximum power level, respectively. All the destinations (D,E,F) will receive all of these three packets. Then, according to the sending and receiving power levels of these packets, node D can calculate the channel information: $h_{A \to D}$, $h_{B \to D}$ and $h_{C \to D}$ with equation (4); node E can calculate the channel information: $h_{A \to E}$, $h_{B \to E}$, $h_{C \to E}$; node F can calculate the channel information: $h_{A \to F}$, $h_{B \to F}$, $h_{C \to F}$. At the beginning of second slot, the receiving nodes D, E, F will send a CSI packet respectively with the channel information they get. For example, D's CSI packet would be sent to A with the channel information $h_{A\to D}$, $h_{B\to D}$, $h_{C\to D}$ and these packets can be overheard by B, C. Nodes E, F would send similar packets to their destination B, C and can be overheard by other neighbors. At the end of this slot, A, B, C can set up a channel state matrix H with the CSI packets they collected. Each of the source node would have a channel state matrix H as equation (5).

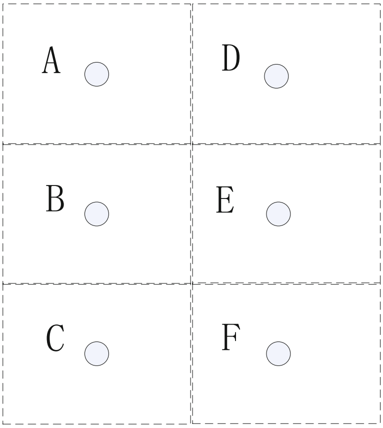


Fig. 4. Six-node topology

$$H = \begin{bmatrix} h_{AD} & h_{AE} & h_{AF} \\ h_{BD} & h_{BE} & h_{BF} \\ h_{CD} & h_{CE} & h_{CF} \end{bmatrix} \tag{5}$$

Then each of the sending nodes would use power control algorithm to allocate their transmission power on DATA packet distributedly. Finally, if the receiving nodes can receive DATA packet correctly, they will send ACK packet with the same power as the DATA packet they received.

3.2 Protocol Details

In the previous subsection, we have given a big picture of UPC-MAC. In this subsection, we will discuss the detailed design of UPC-MAC, including power control algorithm and special channel information matrix.

Power Control Algorithm. Once a node gets the channel state matrix H, it can use a power control algorithm to allocate its transmission power for the DATA packet. Here we use a game theory based algorithm as the basic idea to allocate the transmission power [14] . Our goal is to maximize the overall networks throughput while preventing nodes from using unnecessarily high power. So there are two part in the utility function. The first term denotes the throughput of a link. The achievable throughput is approximated by Shannon capacity theory. The second part reflects the transmission power on this link. To maximize the overall networks throughput, we define a utility function for the power control channel access game. For each link, the utility function is defined as equation (6) :

$$u_i(p_i, \mathbf{p_{-i}}) = log(1 + SINR_i) - \alpha_i \cdot p_i \tag{6}$$

p_i denotes transmission power p on link i; $\mathbf{p_{-i}}$ denotes transmission powers on all the other links except link i; α_i is a positive factor. For example, for link $A \to D$, the utility function is :

$$u_{AD}(p_{AD}, \mathbf{p_{-AD}}) = log(1 + \frac{h_{AD} \cdot P_{AD}}{\sum_{j=1}^{N-1} h_j \cdot p_j}) - \alpha_A \cdot p_{AD} \tag{7}$$

Therefore the optimization problem can be formulated as (8)-(9):

$$max\ u_i(p_i);\ for\ all\ i\ (1 \leq i \leq N) \tag{8}$$

$$\begin{aligned} &s.t. \\ &p_i \in S_i = [0, P_{max}]\ (1 \leq i \leq N) \\ &SINR_i \geq SINR_{threshold}\ (1 \leq i \leq N) \end{aligned} \tag{9}$$

The optimization problem can be solved by game theory, and the solution, if feasible, is the one that achieves Nash Equilibrium (NE). Although NE does not always exist, delightedly we can prove that it does exist in our problem. [15] states that NE exists only if the following 2 conditions are satisfied:

(1) S_i , the set of transmission power for each sender, is a nonempty and convex subset of some Euclidean space.

(2) u_i , the utility function for each sender, is a continuous and quasi-concave function for independent variable p_i

The first condition is readily satisfied. To prove the second condition is also satisfied, we take the second-order derivation of u_i :

$$\frac{\partial^2 u_i}{\partial p_i^2} = -\frac{h_i^2}{(h_i \cdot p_i + \sum_{j=1}^{N-1} h_j \cdot p_j + \sigma^2)^2} \tag{10}$$

Since $\frac{\partial^2 u_i}{\partial p_i^2} < 0$, the second condition is satisfied [15]. Therefore NE exists.

The transmission power of each sending node is defined as the players response function as equation (11). The best response of each node is the transmission power that maximizes its utility function and satisfies C_1 and C_2 .

$$\frac{\partial u_i}{\partial p_i} = -\frac{h_i}{(h_i \cdot p_i + \sum_{j=1}^{N-1} h_j \cdot p_j + \sigma^2)} - \alpha = 0 \tag{11}$$

Solving Equation (12) gives:

$$p_i = \frac{1}{\alpha_i} - \frac{\sum_{j=1}^{N-1} h_j \cdot p_j + \sigma^2}{h_i} \tag{12}$$

In order to guarantee $p_i \leq P_{max}$, we choose $\alpha_i = \frac{1}{P_{max}}$

By rearranging the terms in equation (12), we get a new expression (13)

$$p_i \cdot h_i + \sum_{j=1}^{N-1} h_j \cdot p_j = \frac{h_i}{\alpha_i} - \sigma^2 \tag{13}$$

For all the sending nodes, equation (13) can be denoted as the following matrix form expression:

$$\mathbf{H} \cdot \mathbf{P}^* = \mathbf{G} \tag{14}$$

$\mathbf{H}$ is the $N \times N$ channel state matrix; $\mathbf{P}^*$ is the unique NE solution; $G = [g_1, g_2 ... g_n]^T$ is an $N \times 1$ vector, where $g_i = \frac{h_i}{\alpha_i} - \sigma^2$

So $\mathbf{P}^*$ can be solved by:

$$\mathbf{P}^* = \mathbf{H}^{-1} \cdot \mathbf{G} \tag{15}$$

If the computed p_i does not satisfied the constrain C_1 or C_2, it means that the transmissions cannot be conducted concurrently. In that case, the protocol would deny the sending request with the lowest sending priority and rerun the power allocation algorithm.

Special Channel Information Matrix. As an example, in Fig. 4, nodes A and B want to send a packet to D at the same time. Then node D would receive $RTS_{A \to D}$ and $RTS_{B \to D}$ at the same slot. After receiving these two packets, D will first check priority of the packet and choose the packet with a higher priority, say A as the potential sender. Then it will set $h_{B \to D} = -1$. If $RTS_{A \to D}$ and $RTS_{B \to D}$ have the same priority, it will choose the one that comes earlier.

Once a node receives the matrix with -1 as in equation (16), it knows that it is denied sending this time. Then the matrix can be transfered to a 2×2 matrix as equation (17).

$$H = \begin{bmatrix} h_{AD} & h_{AF} \\ -1 & h_{BF} \\ h_{CD} & h_{CF} \end{bmatrix} \tag{16}$$

$$H = \begin{bmatrix} h_{AD} & h_{AF} \\ h_{CD} & h_{CF} \end{bmatrix} \tag{17}$$

3.3 Discussion

Control Packet Collision Problem. Since in UPC-MAC, two or more senders may send their control packets with the maximum power at the beginning of some slots, these control packets are likely to cause collision at certain destinations. In our protocol design, we do no implement any method to reduce or prevent this kind of collision. But we can prove that the probability of this collision is relatively low.

For a general scenario, we set the parameters as follows: $max\_transmission\_range = 2500m$, $sound\_speed = 1500m/s$, $CSI\_packet\_length = 20B$, $data\_rate = 3kbps$, $guard\_time = 100ms$. So the slot length is 2103ms. We also assume that the distance d between nodes follows uniform distribution $d \sim U[500, 2000]$. In order to avoid collisions, the difference of distance: Δd from two senders to the same receiver should make sure that two control packets will not overlap. That is to say that the difference of the propagation delay for two control packets should be larger than the transmission time of one control packet. Namely, $\frac{\Delta d}{1500m/s} > 20B \times 8/3000bps$.

Based on Geometric probability theory, the maximum probability of collisions is the ratio of the possible collision area to the whole area. With all the parameters above, the maximum collision probability of control packet is approximately 6.3% for two pairs of concurrent communication.

4 Simulation Analysis

In this section, we evaluate the performance of UPC-MAC protocol and compare it with Slotted FAMA. The performance metrics are the network goodput which is measured as the number of successful data transmission per unit time, the relative frequency of concurrent communication which is measured as the ratio between the number of concurrent communications and the number of total transmissions. We conduct simulations using Aqua-sim, an NS-2 based simulator for underwater acoustic networks [16]. We modified the design of Aqua-Sim so that we can set per packet transmission power and range. The simulation parameters are listed in Table 1. These parameters are in accordance with the hardware specifications of underwater OFDM modems.

Table 1. Simulation parameters

DATA Packet Size	100b-500b
DATA Rate	3k bps
Maximum transmission range	1500m
Simulation time	$10^4 s$

4.1 One Hop Network Simulation

In the one-hop network scenario, every node is within the maximum transmission ranges of all the other nodes in the network. In other words, all the exchanged control packets sent with the maximum power can be received by each terminal. We model this scenario in our simulation as follows. Six sensors are deployed in an area by $2000m \times 3000m$ and this area is divided into six square areas with $1000m \times 1000m$. Each sensor node is deployed in the middle of each square field as shown in Fig. 4. Nodes A, B, C are source nodes and nodes D, E, F are

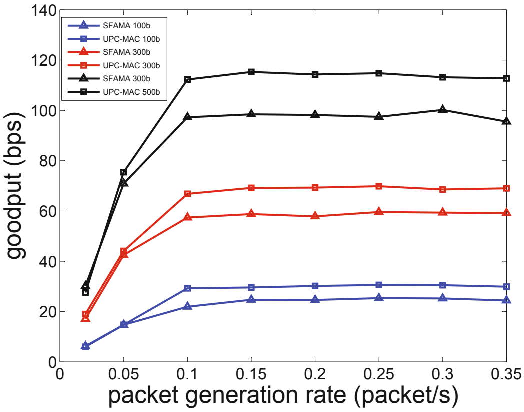
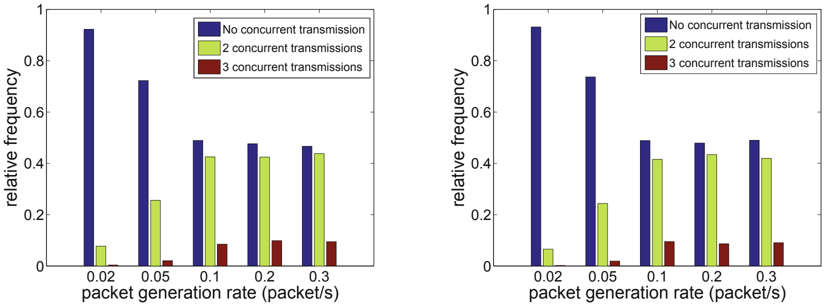


Fig. 5. One Hop Goodput



(a) Packet Length 300b

(b) Packet Length 500b

Fig. 6. Relatively Frequency of Concurrent Communications(One hop)

destinations. Each source node generates packets according to a Possion process with a rate λ. For each generated packet, the destination is randomly selected.

Fig. 5 depicts the network goodput with varying the packet generation rate for the two target MAC protocols, Slotted FAMA and UPC-MAC. This figure shows that UPC-MAC achieves about 15% improvement in network goodput over Slotted FAMA when both of them reach the maximum goodput at a 500b packet length. For the shorter packet lengths, UPC-MAC can achieve about 20% and 17% improvement for 100b and 300b packet length respectively. This improvement comes from the concurrent communications. By contrast, Slotted FAMA does not enable any concurrent communication. The frequency of concurrent communications mainly depends on the network topology. Therefore, different packet lengths yield basically the same frequency of concurrent communications, as shown in Fig. 6(a) and Fig. 6(b) . The reason that packet length has little impact on the frequency of concurrent communication is that both 300b packet and 500b packet can be sent in one slot according to the data rate of the OFDM modem. Therefore the relative frequency of concurrent communication only depends on the positions of senders and frequency of sending packets from these senders. However, the packet length does affect the goodput, as shown in Fig. 5, where a larger packet length leads to a larger goodput.

From Fig. 6(a) and Fig. 6(b) we can also conclude that when the packet generation rate reach $0.1 packet/s$, the relative frequency of concurrent transmissions becomes stable. It is because this is the maximum number of concurrent transmissions. At this situation, the number of transmission can not be increased by adjust the power level. Thus, the goodput curve becomes flat when the packet generation rate reach $0.1 packet/s$. Due to the same reason, the curves in Fig. 5 have the same trend.

4.2 Two-Hop Network Simulation

In this section, we evaluate protocols in a two-hop network, which is a more generic scenario for UWSNs. As shown in Fig. 7, nine nodes are deployed in

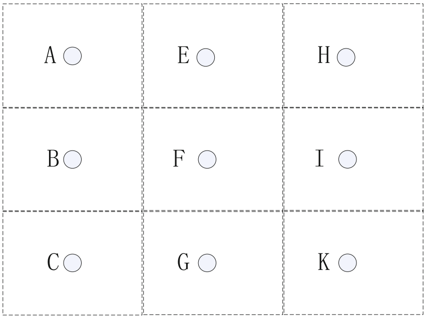


Fig. 7. Nine Nodes Topology

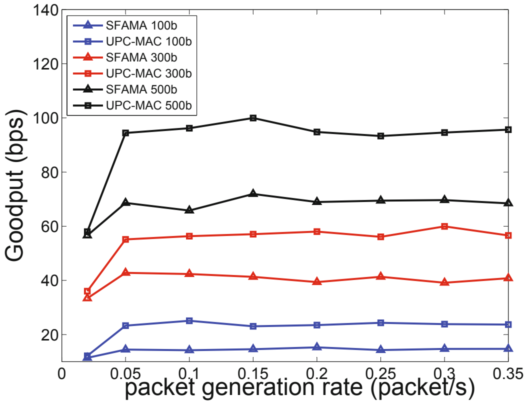
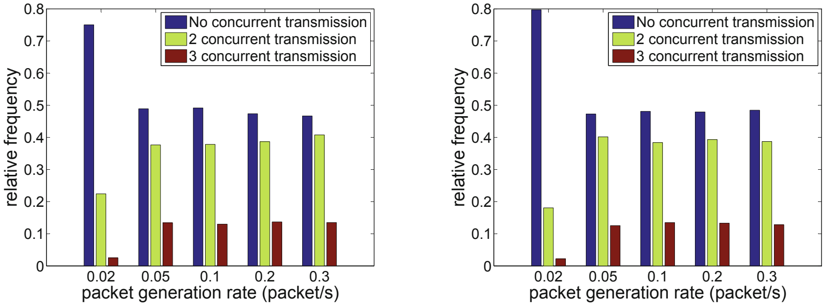


Fig. 8. Two Hop Goodput

a $3000m * 3000m$ area. Nodes A-C are source nodes and H-K are destinations. For each generated packet, the destination is randomly selected. All the other settings are the same with the one-hop scenario. Fig. 8 shows that UPC-MAC can achieve 60%, 45%, 35% improvement in network goodput over Slotted FAMA for three different packet length (100b, 300b and 500b) respectively. The reason for the goodput improvement is the concurrent communication, same as in the one-hop scenario, whose frequencies with different packet lengths are shown in Fig. 9(a) and Fig. 9(b). They yield a similar trend as the one hop network.

By comparing Fig. 5 and Fig. 8, the goodput of the two-hop network shows a larger gap between UPC-MAC and slotted FAMA. The reason is that in this two-hop network, there are more requests for concurrent communications. However, slotted FAMA, for avoiding collisions, can not satisfy such requests. Therefore the goodput of slotted FAMA in such scenario is highly decreased. On the other hand, UPC-MAC, by adjusting the transmission power, fully exploits it's advantages.



(a) Packet Length 300b

(b) Packet Length 500b

Fig. 9. Relatively Frequency of Concurrent Communications(Two hops)

5 Conclusions and Future Work

In this paper, we first introduce the necessity of power control in UWSNs and then propose an underwater power control mac protocol UPC-MAC to improve the goodput in UWSNs. UPC-MAC makes use of the long propagation delay in underwater to collect sending request and channel information. All the nodes can allocate their transmission power distributedly by running a game-theory based power control algorithm. This algorithm enables multiple concurrent transmissions. The simulation shows that UPC-MAC improve the network goodput both in one hop and two hops networks.

Regarding future works, on one hand, since the throughput of UWSNs is highly dependent on the topology and UPC-MAC protocol's performance varies with different topologies, a joint design of deployment and power control channel access algorithm is necessary . On the other hand, transmission rate is another factor that affects the performance of a network and it is highly dependent on transmission power. I would be an interesting direction to model the relationship between transmission rate and transmission power in underwater environments is another direction for future research.

References

1. Cui, J.-H., Kong, J., Gerla, M., Zhou, S.: Challenges: building scalable mobile underwater wireless sensor networks for aquatic applications. IEEE Network, Special Issue on Wireless Sensor Networking 20(3), 12–18 (2006)
2. Akyildiz, I.F., Pompili, D., Melodia, T.: Underwater acoustic sensor networks: Research challenges. Ad Hoc Networks 3(3), 257–279 (2005)
3. Chitre, M., Shahabudeen, S., Stojanovic, M.: Underwater acoustic communicatin and networks: Recent advances and future challenges. Marine Technology Society Journal 1, 103–116 (2008)
4. Luo, Y., Pu, L., Peng, Z., Zhou, Z., Cui, J.-H.: CT-MAC: a MAC protocol for underwater MIMO based network uplink communications. In: Proceedings of the Seventh ACM International Conference on Underwater Networks and Systems, p. 23. ACM (2012)
5. Stojanovic, M.: Underwater acoustic communication. Wiley Encyclopedia of Electrical and Electronics Engineering (1999)
6. Karn, P.: Maca-a new channel access method for packet radio. In: ARRL/CRRL Amateur Radio 9th Computer Networking Conference, vol. 140, pp. 134–140 (1990)
7. Bharghavan, V., Demers, A., Shenker, S., Zhang, L.: Macaw: a media access protocol for wireless lan's. In: ACM SIGCOMM Computer Communication Review, vol. 24(4), pp. 212–225. ACM (1994)
8. Garcia-Luna-Aceves, J., Fullmer, C.L.: Floor acquisition multiple access (fama) in single-channel wireless networks. Mobile Networks and Applications 4(3), 157–174 (1999)
9. Molins, M., Stojanovic, M.: Slotted fama: a mac protocol for underwater acoustic networks. In: OCEANS 2006-Asia Pacific, pp. 1–7. IEEE (2007)
10. Partan, J., Kurose, J., Levine, B.N., Preisig, J.: Spatial reuse in underwater acoustic networks using rts/cts mac protocols. University of Massachusetts Dept. of Computer Science, UM-CS-2010-045 (2010)
11. Baccelli, F., Blaszczyszyn, B., Mühlethaler, P., et al.: A spatial reuse aloha mac protocol for multihop wireless mobile networks (2003)
12. Xu, K., Gerla, M., Bae, S.: Effectiveness of rts/cts handshake in ieee 802.11 based ad hoc networks. Ad Hoc Networks 1(1), 107–123 (2003)
13. Rappaport, T.S., et al.: Wireless communications: principles and practice, vol. 2. Prentice Hall PTR, New Jersey (1996)
14. Wang, F., Younis, O., Krunz, M.: Gmac: A game-theoretic mac protocol for mobile ad hoc networks. In: 2006 4th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, pp. 1–9. IEEE (2006)
15. Saraydar, C.U., Mandayam, N.B., Goodman, D.J.: Efficient power control via pricing in wireless data networks. IEEE Transactions on Communications 50(2), 291–303 (2002)
16. Xie, P., Zhou, Z., Peng, Z., Yan, H., Hu, T., Cui, J.-H., Shi, Z., Fei, Y., Zhou, S.: Aqua-sim: an ns-2 based simulator for underwater sensor networks. In: OCEANS 2009, MTS/IEEE Biloxi-Marine Technology for Our Future: Global and Local Challenges, pp. 1–7. IEEE (2009)

FMAC for Coexisting Ad Hoc Cognitive Radio Networks

Yanxiao Zhao[1], Min Song[2], and ChunSheng Xin[3]

[1] Department of Electrical and Computer Engineering
South Dakota School of Mines and Technology, Rapid City, SD 57701 USA
yanxiao.zhao@sdsmt.edu
[2] Electrical Engineering and Computer Science Department
The University of Toledo, Toledo OH 43606 USA
min.song@utoledo.edu
[3] Department of Electrical and Computer Engineering
Old Dominion University, Norfolk 23508 USA
cxin@ieee.org

Abstract. Media access control plays a critical role in cognitive radio networks (CRNs). In our previous work, we have proposed a *fairness-oriented media access control* (FMAC) protocol to achieve fair and efficient coexistence of infrastructure-based CRNs. In this paper, we enhance FMAC to be used for coexisting ad hoc CRNs, where no centralized base stations exist, and *secondary users* (SUs) access channel independently. In FMAC, the contention window size is essential to network performance such as throughput. We first derive the optimal contention window size, which can then be used by SUs to achieve optimal throughput. However, the optimal contention window size is closely related to the total number of users of all CRNs, which is typically unknown to each individual SU of coexisting ad hoc CRNs. We attack this problem by building a bridge between the *average number of consecutive idle time slots* and *optimal contention window size*, since the average number of consecutive idle time slots can be easily observed by each individual SU. Hence, SUs can independently adjust their contention window size by observing their current average number of consecutive idle time slots and eventually approach the optimal contention window without the information of the total number of SUs. Extensive simulations are conducted and the results verify that the enhanced FMAC is able to significantly improve the fairness among coexisting ad hoc CRNs while maintaining good throughput.

Keywords: coexisting, cognitive radio networks, fairness, throughput.

1 Introduction

The rapid proliferation of various wireless applications results in the issue known as *spectrum scarcity*, which is gaining intensified attention by authorities, industries and academia. *Cognitive radio network* (CRN) is commonly viewed as a

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 391–401, 2013.

disruptive technology to relieve this issue and significantly improve spectrum efficiency. In recent years, CRNs have attracted considerable attention and related research on various topics can be found in the literature [1–4].

With cognitive radio being a fundamental technology for future wireless communications, we expect that CRNs will be ubiquitous as today's highly congested WiFi networks. *Media access control* (MAC) protocol design is of fundamental importance for future ubiquitous coexisting CRNs. In this paper, we present a MAC for coexisting *ad hoc CRNs*. In ad hoc CRNs, there are no centralized devices such as base stations to coordinate SUs. SUs independently sense the channel state and access the channel once it is detected idle. In such networks, MAC protocol design plays a critical role. A variety of MAC protocols have been proposed for CRNs, which are generally classified into three categories: random access, time slotted, and hybrid. Due to the difficulty of time synchronization in ad hoc networks, it is not practical to adopt time slotted protocols [5]. Instead, the random access based approach is viewed promising, of which *carrier sense multiple access* (CSMA) based protocols are popularly proposed [6–9].

For the ubiquitous coexisting CRNs, fairness among SUs is a major concern but has not yet received sufficient attention in the literature. Fairness measures the ability of SUs to share a common channel equally. Two well-known fairness schemes are bandwidth-based fairness and time-based fairness. In this paper, we adopt the time-based fairness scheme. For instance, considering N users competing for the same channel, the ideal fairness is achieved if each user is assigned with $1/N$ time over the total period observed.

In the existing CRN MACs, the two-state sensing model is commonly assumed, which classifies a channel into only two states: *idle* or *busy*. With the rapid proliferation of wireless services, the number of SUs typically exceeds the number of channels. As a result, it is very likely that multiple SUs have to compete for the same channel. If a MAC protocol based on the two-state sensing model is used for coexisting ad hoc CRNs, then when one SU is accessing the channel, SUs from other networks would starve since they falsely perceive that the channel is being used by a *primary user* (PU). Hence this can result in poor fairness among SUs. Motivated by this issue, we propose a fairness-oriented MAC, with which SUs are able to share a channel friendly.

In our previous work [10], we have proposed a novel MAC protocol for CRNs, termed *fairness-oriented media access control* (FMAC), to achieve fair and efficient coexistence of infrastructure-based CRNs. Different from the existing MAC protocols in CRNs, FMAC is designed using a three-state sensing model, which classifies a channel into three states: H_0 *(idle)*, H_1 *(occupied by a PU)*, and H_2 *(occupied by an SU)*. SUs respond differently depending on the channel state. Specifically, with FMAC, when the channel is detected as busy, an SU does not simply switch to a new channel. Instead, it utilizes a spectrum sensing algorithm to further determine whether the channel is being used by a PU or another SU. In the latter case, the sensing SU may choose to compete for channel access with the SU that is accessing the channel. In summary, as long as a channel is not used by the PU, all SUs can fairly compete for channel access and achieve maximum fairness among them.

In this paper, we enhance FMAC for coexisting ad hoc CRNs which raise new changes due to lack of centralized base stations to obtain critical information such as the number of SUs. For FMAC, the size of contention window is an essential factor to affect network performance such as the successful access probability and throughput. As such, we derive the optimal contention window size to maximize the success probability of accessing a channel, which then optimizes throughput as well. However, we will see that the optimal size of contention window is closely related to the total number of users in the entire network, while the total number of SUs is unknown to each individual SU in ad hoc CRNs, due to lack of base stations. We solve this problem by building a 'bridge' between the *average number of consecutive idle time slots* and *the optimal window size*, since the average number of consecutive idle time slots is observable by each individual SU. Hence, each SU can estimate the optimal window size from the observed average number of consecutive idle time slots, and independently adjusts its contention window size without the information of the number of SUs. This research also sheds light on optimal performance analysis of FMAC-based CRNs.

The main contributions of this paper are summarized as follows.

- We have enhanced FMAC for fair and efficient coexistence of ad hoc CRNs.
- We have derived the optimal contention window size to obtain optimal throughput.
- We have proposed a novel approach to build a bridge between the optimal contention window size and the average number of consecutive idle time slots, so that optimal throughput can be achieved without the information of the number of SUs, which is often difficult to be obtained for ad hoc CRNs.

The rest of the paper is organized as follows. Section 2 describes the fairness-oriented MAC protocol for coexisting ad hoc CRNs. Section 3 presents the theoretical analysis of FMAC. Simulation results are presented in Section 4. Concluding remarks are drawn in Section 5.

2 Fairness-Oriented MAC Design

We focus on coexisting ad hoc CRNs, in which one PU and multiple SUs that are associated with different CRNs exist. All SUs independently sense and access the channel. The channel state is classified into three types: H_0 (*idle*), H_1 (*occupied by a PU*) and H_2 (*occupied by an SU*), illustrated as follows [11]:

$$r_i = \begin{cases} n_i, & H_0 \\ x_p + n_i, & H_1 \\ x_s + n_i, & H_2 \end{cases}, \qquad (1)$$

where x_s is the signal that an SU transmits, x_p is the signal that the PU transmits, r_i is the signal that an SU received, n_i is the zero-mean additive white Gaussian noise (AWGN).

Based on the three-state sensing model, we have proposed FMAC for coexisting infrastructure-based CRNs in [10]. The main idea is briefly described in

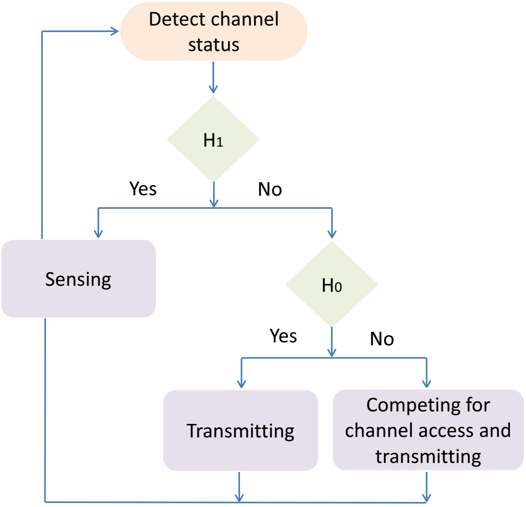


Fig. 1. Flow chart of FMAC

the following. As illustrated in Fig. 1, SUs take distinct actions based on the state of a channel: H_0, H_1, or H_2. Specifically, an SU accesses the channel if the channel state is H_0. If the channel state is H_1, the SU keeps silent and continues to monitor the channel's state. If the channel state is H_2, the SU knows that the channel is being used by an SU instead of the PU. Therefore, it can participate in competition for channel access. If an SU is transmitting, other SUs keep monitoring the channel status. An SU has to vacate the channel whenever the PU appears.

Next we describe how multiple SUs compete for the same channel under the state of H_2. An SU monitors the channel activity when it has a packet to transmit. The SU starts transmitting only when an idle period equals a Distributed Inter-Frame Space (DIFS). In the case that the channel is busy, the SU randomly selects a backoff interval from (0, W-1), where W represents the size of the contention window. The backoff time counter is decremented whenever the channel is sensed idle, stopped when a transmission is detected and reactivated when the channel is sensed idle again for the duration of a DIFS. Finally, the SU transmits when the backoff time counter reaches 0. In the case that a collision occurs, the above precess is repeated.

Note that in FMAC, a constant contention window size W is employed for all SUs, which results in the same access probability among competing SUs and thus leads to the optimal fairness. The derivation of an appropriate W is deferred to Section 3.1.

3 Performance Analysis

We consider two crucial performance metrics: throughput and fairness. The throughput and fairness for FMAC in coexisting infrastructure-based CRNs have been analyzed in [10]. In this paper, we focus to address a new challenge raised by ad hoc CRNs, i.e., how to achieve optimal throughput without critical information such as the number of SUs. First, we derive the optimal contention window size. To be seen shortly, the optimal contention window size depends on the total number of SUs. Unfortunately, the information of number of SUs is not known to each individual SU in ad hoc CRNs. Therefore, each individual SU is hard to use this parameter to achieve optimal throughput. One of the novel contributions of this work is to find an alternate parameter that is observable at an individual SU, while closely related to the optimal window size. We have found that the parameter satisfying this requirement is the *average number of consecutive idle time slots*. We then build a 'bridge' between the *average consecutive idle time slots* and *the optimal window size*. Therefore, SUs can estimate the optimal contention window size from their observed average number of consecutive idle time slots; hence SUs can independently adjust their contention window size to approach the optimal contention window without knowing the information of the number of SUs. The detailed analysis is presented in the following subsections.

3.1 Optimal Window Size

Let N denote the total number of SUs in coexisting ad hoc CRNs. We virtually divide the time into slots based on the channel activity. For a given time slot, the probability that the PU is active or inactive is denoted as P_1 and P_0, respectively. Therefore, a successful SU transmission occurs with a probability P_s, which can be derived as:

$$P_s = N \cdot P_a(1 - P_a)^{(N-1)} \cdot P_0, \tag{2}$$

where P_a denote the channel access probability and has been theoretically derived in [10]. Correspondingly, the consecutive idle probability of a given time slot is:

$$P_i = (1 - P_a)^N \cdot P_0. \tag{3}$$

Let us examine the relationship between the optimal window size W and the total number of users N. Based on the result of [10], we approximate P_a as $P_a = \gamma/W$, where $0 < \gamma \leq 1$. γ can be viewed as a constant coefficient and will be examined by simulations. Substituting $P_a = \gamma/W$ into Eq. 2, then P_s can be re-written as:

$$P_s = N \cdot P_0 \cdot \frac{\gamma}{W}(1 - \frac{\gamma}{W})^{N-1}. \tag{4}$$

Motivated by the intuition that a higher successful probability leads to a higher throughput, we can find the optimal contention window, denoted by W^*, to achieve the maximum P_s. Specifically, let

$$\frac{dP_s}{dW} = 0.$$

After some calculations, we obtain the optimal W, denoted by W^*, as follows:

$$W^* = \gamma N. \tag{5}$$

From Eq. 5, we can see that W^* depends on the number of SUs in the system. Note that all SUs have to adopt the same W^* in a distributed way. However, a great challenge is that the total number of SUs is unknown to an individual SU; hence Eq. 5 cannot be directly used by SUs to achieve optimal throughput. Next we proceed to find an alternate parameter that is closely related to W^*, while observable by an individual SU.

3.2 Average Number of Consecutive Idle Time Slots

To solve the challenge discussed above, we seek to build a 'bridge' between the *average number of consecutive idle time slots* and *the optimal window size.* Since the average number of consecutive idle time slots is observable by individual SUs, an SU can thus estimate the optimal window size from the observed average number of consecutive idle time slots, and hence independently adjusts its contention window size to approach the optimal contention window without knowing the information of the number of SUs.

First, let us study a complete contention backoff cycle, i.e., the period that the backoff timer decreases from $i \in [0, W-1]$ to 0. This period consists of idle, collision, SU transmission, and PU occupation time slots. Next, we examine the consecutive idle slots in this period. It is clear that in one backoff period, there are up to $(W-1)$ consecutive idle slots. In the following, we calculate the average number of consecutive idle slots. From Eq. (3), we can get the idle probability for a given time slot. Therefore the average number of consecutive idle time slots, denoted by $\overline{X}$, can be derived by averaging all the possible situations with consecutive idle slots. For example, the probability of one idle slot is P_i, the probability of two consecutive idle slots is P_i^2, ..., and so on and so forth. Therefore we have

$$\begin{aligned}\overline{X} &= \sum_{k=1}^{W-1} k \cdot P_i^k = P_i \sum_{k=1}^{W-1} k \cdot P_i^{k-1} \\ &= P_i(\sum_{k=1}^{W-1} P_i^k)' = P_i(\frac{P_i - P_i^W}{1-P_i})' \\ &= \frac{[1 - P_i^{W-1} - (W-1)(1-P_i)P_i^{(W-1)}]P_i}{(1-P_i)^2}.\end{aligned} \tag{6}$$

When $W >> 1$, it is reasonable to assume $P_i^{(W-1)} \doteq 0$. So Eq. (6) can be simplified to

$$\overline{X} \approx \frac{P_i}{(1-P_i)^2}. \tag{7}$$

From Eq. (5), we have $N = W^*/\gamma$. Substituting $N = W^*/\gamma$ into Eq. 7 yields

$$\overline{X} \approx P_0(1 - \frac{\gamma}{W})^{W^*/\gamma}[1 - \frac{\gamma}{W})^{W^*/\gamma}]^{-2}, \tag{8}$$

where W is the current contention window size being used by an SU. From Eq. (8), W^* can be estimated using the current observed number of consecutive idle time slots $\overline{X}$. In other words, if we represent Eq. (8) as $\overline{X} = f(W^*)$, then W^* is derived as $W^* = f^{-1}(\overline{X})$. This relationship clearly demonstrates that it is practical to dynamically adjust the contention window size by observing the average number of consecutive idle time slots, and eventually approach the optimal contention window size.

4 Performance Evaluation

In this section, we will first examine the optimal size of the contention window in terms of γ. Next we will evaluate the SU throughput and fairness of FMAC in coexisting ad hoc CRNs. We compare FMAC with another scheme, in which the IEEE 802.11 is used together with the two-state sensing model, termed *Two-state MAC* (TMAC). We will demonstrate that the fairness of FMAC is dramatically improved, while maintaining a similar throughput as TMAC.

4.1 Throughput

In the first experiment, we examine the optimal size of the contention window. The PU's activity is modeled as $P_0 = P_1 = 0.5$. Each CRN consists of 20 SUs and 3 coexisting CRNs are studied. With TMAC, the channel access competition occurs among SUs within one CRN at a time, while with FMAC, all SUs from all CRNs participate in the competition. We set $\gamma = 0.5$. The throughput and the successful probability of channel access are shown in Fig. 2.

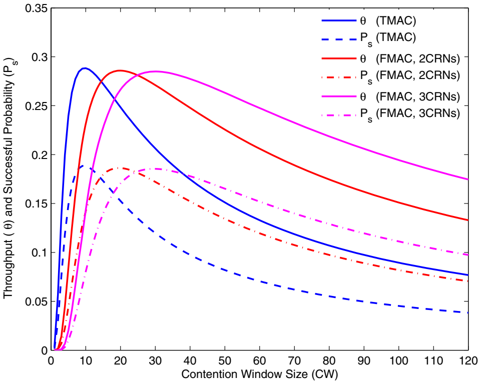


Fig. 2. Throughput and successful probability of channel access (P_0=0.5, $\gamma = 0.5$)

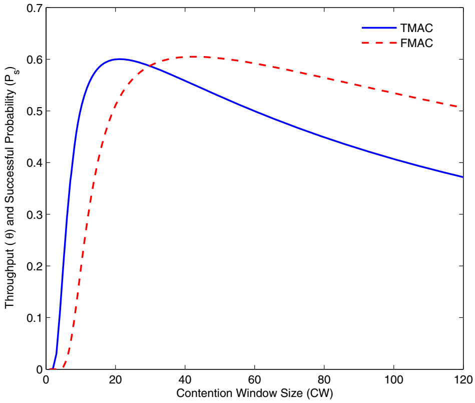


Fig. 3. Throughput vs. window size with two coexisting CRNs (P_0=0.8,$\gamma = 1$)

From the simulation results, we find that both the throughput θ and successful probability of channel access P_s are maximized at the same window size. This verifies that the derived optimal size of contention window in terms of P_s also applies to the optimal throughput. As a matter of fact, we further discover that the optimal throughput (peak value) is reached at $W^* = 0.5N$, which matches our analysis result (Eq. 5) presented in Section 3.1. Furthermore, we notice that the optimal window size is larger in FMAC than in TMAC. This is because more SUs are involved in the competition of channel access. Next, we examine the scenario with $\gamma = 1, P_0 = 0.8$ and two CRNs: CRN1 and CRN2. CRN1 has 15 SUs and CRN2 has 25 SUs, respectively. Fig. 3 shows the throughput of FMAC and TMAC with distinct contention window sizes. A similar trend can be found with Fig. 2, but with a higher throughput due to $P_0 = 0.8$.

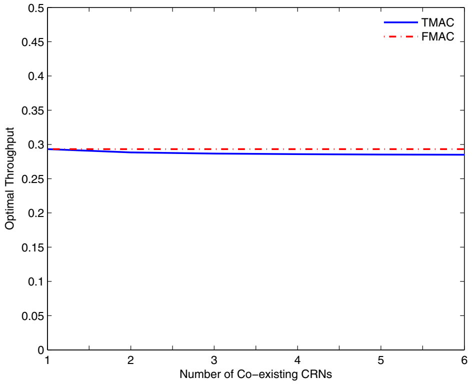


Fig. 4. Optimal throughput of FMAC and TMAC with coexisting CRNs, P_0=0.5

Fig. 4 plots the optimal throughput of TMAC and FMAC with multiple CRNs. The optimal window size used to obtain the throughput is obtained by observing the average number of consecutive idle time slots, as discussed in Section 3.2. Each CRN has 10 SUs. With TMAC, only 10 SUs access the channel and other SUs keep silent. In contrast, with FMAC, more SUs equally share and compete for the channel, so the competition among SUs is high, and may degrade the throughput. However, Fig. 4 indicates that the optimal throughput of FMAC is almost the same as the one of TMAC. In other words, FMAC maintains a similar throughput as TMAC.

Next, we examine the impact of the PU's activity on SU throughput, as plotted in Fig. 5. Each CRN has 5 SUs. It can be seen that the optimal throughput of SUs gradually improves when P_0 increases. This is because SUs have a better chance to transmit when the PU occupies the channel for less time.

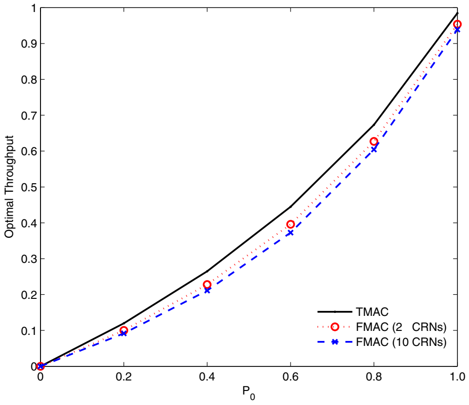


Fig. 5. Optimal throughput of FMAC and TMAC as a function of the PU activity (P_0)

4.2 Fairness

Fairness performance is evaluated in this subsection. We assume two CRNs coexist, with each CRN having 5 SUs. The observation time is 20 periods. The transmission time of each SU is recorded continuously in these 20 periods. We calculate the accumulative fairness in the first k (k=1,2,...,20) periods using Jain Index. It is known that a higher Jain Index implies better fairness performance. In the case of FMAC, all SUs of all CRNs contend for channel access with an equal access probability P_a. However, with TMAC, one CRN will continuously occupy the channel once it wins the competition. To maintain a good fairness level among the SUs of the same CRN, we assume each SU can only transmit 10 successive frames once it wins the channel. Fig. 6 shows the Jain Index for 20 periods with two and four coexisting CRNs. It can be seen that the fairness of FMAC is dramatically improved compared with TMAC.

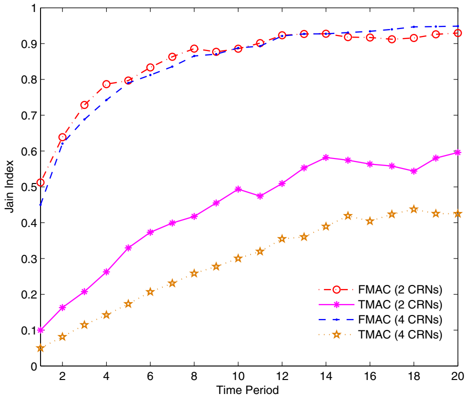


Fig. 6. Fairness comparison between FMAC and TMAC

5 Conclusions

In this paper, we have enhanced the fairness-oriented media access control (FMAC) protocol for coexisting ad hoc CRNs. We have derived the optimal window size to maximize the successful channel access probability and hence the throughput. In coexisting ad hoc CRNs, an individual SU has no clue about the total number of SUs, which makes it hard for each SU to select the optimal contention window size to obtain optimal throughput. We have solved this problem by building a ‘bridge’ between the average number of consecutive idle time slots and the optimal window size, since the average idle time slots can be easily observed by each individual SU. Hence, SUs can independently adjust their contention window size by observing the current average number of consecutive idle time slots to eventually approach the optimal contention window without knowing the information of the number of SUs in the system. Simulation results have verified that FMAC is able to significantly improve the fairness while maintaining a good throughput.

Acknowledgement. The research of Min Song is supported in part by NSF grant CNS-1248092 and NSF IPA Independent Research and Development (IR/D) Program. However, any opinion, finding, and conclusions or recommendations expressed in this material; are those of the author and do not necessarily reflect the views of the National Science Foundation. The research of ChunSheng Xin is supported in part by NSF under grants CNS-1217668, ECCS-1247853, and CNS-1017172.

References

1. Gao, L., Wang, X., Xu, Y., Zhang, Q.: Spectrum trading in cognitive radio networks: A contract-theoretic modeling approach. IEEE Journal on Selected Areas in Communications 29(4), 843–855 (2011)

2. Qiu, R., Hu, Z., Li, H., Wicks, M.: Cognitive radio network. Cognitive Radio Communications and Networking: Principles and Practice, 381–426 (2012)
3. Akyildiz, I., Lo, B., Balakrishnan, R.: Cooperative spectrum sensing in cognitive radio networks: A survey. Physical Communication 4(1), 40–62 (2011)
4. Zhao, Y., Song, M., Xin, C.: A weighted cooperative spectrum sensing framework for infrastructure-based cognitive radio networks. Computer Communications 34(12), 1510–1517 (2011)
5. Cormio, C., Chowdhury, K.R.: A survey on mac protocols for cognitive radio networks. Ad Hoc Networks 7(7), 1315–1329 (2009)
6. Chong, J.W., Sung, Y., Sung, D.K.: Rawpeach: Multiband csma/ca-based cognitive radio networks. Journal of Communications and Networks 11 (2009)
7. Thoppian, M., Venkatesan, S., Prakash, R.: Csma-based mac protocol for cognitive radio networks. In: IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1–8 (2007)
8. Chen, R.R., Liu, X.: Coexisting with csma-based reactive primary users. In: IEEE Symposium on New Frontiers in Dynamic Spectrum, pp. 1–7 (2010)
9. Akyildiz, I., Lee, W.Y., Chowdhury, K.: Spectrum management in cognitive radio ad hoc networks. IEEE Network 23(4), 6–12 (2009)
10. Zhao, Y., Song, M., Xin, C.: Fmac: A fair mac protocol for coexisting cognitive radio networks. In: IEEE INFOCOM (2013)
11. Zhao, Y., Song, M., Xin, C., Wadhwa, M.: Spectrum sensing based on three-state model to accomplish all-level fairness for co-existing multiple cognitive radio networks. In: IEEE INFOCOM, pp. 1782–1790 (2012)

Effective RSS Sampling for Forensic Wireless Localization

Yinjie Chen[1], Zhongli Liu[1], Xinwen Fu[1], and Wei Zhao[2]

[1] Computer Science Department, University of Massachusetts Lowell
{ychen1,zliu,xinwenfu}@cs.uml.edu
[2] University of Macau
weizhao@umac.mo

Abstract. In many applications such as wireless crime scene investigation, we want to use a single device moving along a route for accurate and efficient localization without the help of any positioning infrastructure or trained signal strength map. Our experiments show that in a complicated environment, such as building corridors and downtown areas, triangulation or trilateration cannot be used for accurate localization via single device. A simple approach, which is better and robust, is to use where the maximum RSS (received signal strength) is sensed as the target's location. The question is how to make sure the maximum RSS is received while moving. Our novel RSS sampling theory presented in this paper answers this question: if RSS samples can reconstruct a target transmitter's power distribution over space, the location corresponding to the peak of such power distribution is the target's location. We apply the Nyquist sampling theory to the RSS sampling process, and derive a mathematical model to determine the RSS sampling rate given the target's distance and its packet transmission rate. To validate our RSS sampling theory, we developed BotLoc, which is a programmable and self-coordinated robot armed with a wireless sniffer. We conducted extensive simulations and real-world experiments and the experimental results match the theory very well. A video of BotLoc is at `http://youtu.be/FsWLrH8Nj50`.

1 Introduction

With the booming WiFi networks comes wireless network based crimes. One example explains this type of cyber crime scene well. According to the Huffington Post news on September 24, 2011 [30], law enforcement stormed into a Buffalo homeowner at 6:20 A.M, accusing him of sharing files for videos and images of children engaged in sexual acts. The fact is that his neighbor was using this unlucky man's open WiFi and committed the crime with a peer-to-peer file sharing software. At the end, the law enforcement released the unlucky man and arrested the real suspect after correlating logs of the involved peer-to-peer file sharing software from the offender and discovering his real ID. The Huffington Post news also gave example of similar crimes and other similar crimes were also reported elsewhere [4, 26, 28, 2, 18]. We also interviewed local state police, who

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 402–413, 2013.

confirmed such crimes have been occurring periodically. The Huffington Post news also cited survey result from the WiFi alliance, revealing that 32 percent of American people used a Wi-Fi network belonging to somebody else. This sheer fact explains why crimes over WiFi networks are widely spread. Further more, tech savvy criminals may also crack encrypted WiFi networks for crimes.

In this paper, we develop wireless network forensics techniques for fighting cyber crimes such as distribution of child pornography and cyber terrorism for public safety and homeland security. We target crime scenes of WiFi networks as discussed above, where people are mistakenly arrested for downloading child pornography. Such awkward moments not only bring much inconvenience to innocent people but also may alert real suspects, who may destroy the evidence to exonerate themselves. Not all investigators may be as lucky as the one in the Huffington Post news, who can use other online logs to identify the real suspect. To avoid such mishaps, an effective option for law enforcement is to use *mobile* and standalone WiFi sniffers, move around in the suspect region (driving a car or walking) and sense wireless signals transmitted from the target and locate its position. Such types of devices should be rapid in response, easy to operate, and have a universal use in various environments. Such a sniffer can be used in a legal way. Use the wireless crime scene investigation as an example. Before raiding the house, investigators can first acquire a search warrant or court order and use the WiFi sniffer to survey the suspect area and locate the real suspect. If necessary, another search warrant can be secured to search the real suspect. Please refer to Section 2 for further discussion of the legal matter.

We conduct localization experiments in a complicated environment, such as building corridors, and observe that, triangulation or trilateration often performs bad localization via single mobile device taking multiple measurements at different locations. Localization in other complicated environments such as downtown areas also produces poor localization accuracy. A simple approach for a single device locating a target is to use where the maximum RSS is sensed as the target's location. Therefore, a question to answer is how to make sure the maximum RSS is received while moving. Astute readers may answer this question instantly: use over-sampling and move slowly to collect as many RSS samples as possible. However, what do we mean by over-sampling? How slow is slow? We will address these fundamental issues in this paper.

We propose the *RSS sampling theory* to address the problem of sampling: if RSS samples can reconstruct a target transmitter's power distribution over space, the location corresponding to the peak of such power distribution is the target's location. Our RSS sampling theory explains the relationship between the sniffer's velocity, target traffic pattern, and the distance between the target and sniffing route.

We implemented a prototype system, *BotLoc*, to demonstrate the usability of our RSS sampling theory. BotLoc is a programmable and self-coordinated robot (P3-DX [14]) armed with a wireless sniffer and it can be used for both indoor and outdoor localization.We introduce a set of localization schemes powered by the RSS sampling theory for BotLoc. With an appropriate velocity, the robot correctly collects RSS samples along its route and is able to derive the maximum RSS at the intersection point of the route and its perpendicular line through the target. Using this method, we can determine the angle of the target with respect to the route.

Our major contributions can be summarized as follows:

- We are the first to propose the RSS sampling theory for a moving sniffer, which can be carried by humans or vehicles, to reconstruct the target power distribution via a sampled RSS time series. We demonstrate the use of this RSS sampling theory via precise localization using a moving robot. We proposed a set of general theories to enhance the accuracy and efficiency of wireless mobile localization in this context. We explicitly formulated the relationship between the robot's velocity, the target wireless device's packet transmission interval, and the distance between the target and robot's route. This RSS sampling theory guarantees sufficient RSS collection for precise localization and avoids over-sampling, or in other words, answers what is over-sampling.
- We developed a fully functional localization system: BotLoc. BotLoc is a P3-DX robot armed with a wireless sniffer. It applies our RSS sampling theory for collecting RSS samples and locating a target mobile device. It is infrastructure-free and training-free. Our contribution also includes extensive experimental results which demonstrate the effectiveness of the proposed theorem.

The rest of this paper is organized as follows. Section 2 introduces our motivation to develop RSS sampling theory for forensic localization in a wireless network crime scene investigation and related legal issues. Section 3 presents the application background, problem definition of RSS sampling and the solution, the RSS sampling theory. Section 4 introduces the BotLoc system for localization and presents experimental evaluation of BotLoc. Section 5 discusses related work, and Section 6 concludes this paper.

2 Motivation and Legal Issues

As we discussed in the introduction, the pervasive deployment of WiFi has provided an easy venue for cyber criminals to commit crimes including accessing illegal content anonymously [4, 26, 28, 30, 2, 18]. Another interesting case is that experienced hackers also utilize public WiFi network to commit crimes. The notorious hacker, Max Butler, who was sentenced to 13 years in prison [22] in February 2010, often stayed at a large hotel in downtown San Francisco and used open wireless networks or hacked wireless networks with weak encryption to commit remote attacks while hiding behind wireless routers.

The difficulty for law enforcement is that, as the network address translation technique (NAT) is widely used in wireless routers, the observed public IP address may not reveal the real location information of the criminals. Therefore, law enforcement is not able to tell wether the criminal is in the house which is associated with the observed IP address, or in a neighbour's house. Moreover, without evidence to prove that the criminal is a neighbour, law enforcement can not get a search warrant to carry out the search in that neighbour's house. To address this problem, Yang *et al.* [32] proposed methods to remotely identify whether a criminal is using wireless or wired network while committing crimes. If the criminal is using wired network to commit crimes, then the criminal should be in the house where the observed public IP address is associated. As the authors acknowledge, the robustness and effectiveness of their approach need further improvement.

To the best of our knowledge, currently there is no effective and efficient technique to locate a criminal which is committing crimes using a wireless network. The best solution is to develop a forensic localization tool for wireless network crime scene investigation. This tool could help law enforcement to collect evidence about criminals' location. From an interview with local state police, we have confirmed that law enforcement urgently needs such a localization toolkit, and requires the toolkit to be portable, infrastructure-free, and not reliant on any training data such as a trained RSS map. Besides, forensic localization using such toolkit should not violate the law.

To conduct forensic localization legally, the localization toolkit should be used in the following way. First of all, the law enforcement identifies illegal activities that are coming from a public IP address which is associated to a house A. Take the peer-to-peer file sharing of child pornography as an example. The investigators download child pornography from a peer and acquires the associated IP address. If the IP address is associated with a broadband connection rather than a cellular connection, an ISP's billing record can be subpoenaed to recover a physical address [32]. Secondly, law enforcement obtains authorization, a wiretap, to monitor the activities of A's router, and derives MAC address of the target machine that is performing illegal activities. Thirdly, the law enforcement uses a forensic localization toolkit to collect only wireless signal strength information which is related to that target machine. This localization requires legal authorization such as a subpoena or a court order. To collect necessary wireless traffic in order to perform the localization, law enforcement can download the illegal content from the suspect computer using the peer-to-peer file sharing software. Finally, after the target computer is located, law enforcement obtains a search warrant and searches the suspect's computer.

In the rest of this paper, we focus on how to collect wireless traffic and locate the suspect.

3 RSS Sampling

We devote this section to the problem of RSS sampling because of its importance in various application scenarios. We will first discuss the application background, then define the problem of RSS sampling and finally present the solution.

3.1 Background

As we mentioned in the introduction, a simple approach for a single device locating a target is to use where the maximum RSS is sensed as the target's location. We will propose the *RSS sampling theory* to address the problem of sampling: if RSS samples can reconstruct a target transmitter's power distribution over space, the location corresponding to the peak of such power distribution is the target's location where we sense the maximum RSS from the target.

Since we are concerned with RSS distribution over space, we now introduce a wireless propagation path loss model in (1) [8, 6]. It gives the relationship between the distance and RSS at a receiver,

$$P(d) = P(1) - 10\alpha \log(d) - W + X\sigma, \tag{1}$$

where distance d (in meters) is the receiver-transmitter distance and power $P(d)$ is the RSS at the receiver's antenna. α is the path loss exponent, W (in dB) is the wall attenuation degree, and $X\sigma$ is a normally distributed variable with mean of 0 and variance of σ^2. X_σ is caused by phenomena including multipath propagation. This log normal wireless propagation model is merely an approximation. We find this approximation is accurate enough to address our problem.

3.2 The Problem

The RSS sampling problem is defined as follows. A moving wireless sniffer, carried by a vehicle or human, moves along a route and collects RSS samples (a RSS map) along a route. The moving velocity is adjustable. How can we sample RSS so that these RSS samples can be used to reconstruct the target's transmission power distribution over the route? This simple scenario catches the essence of the sampling problem and also has its own application, such as surveillance along a corridor.

Without loss of generality, we use a moving robot equipped with a wireless sniffer to explain the RSS sampling process. A robot is used because it is self coordinated and its speed measurement is simple and stable with accompanying robot APIs. Figure 1 shows an example of the target power distribution $s(t)$ over a route. Those dots below the curve $s(t)$ represent RSS samples collected by the robot armed with a sniffer. The target's orthogonal projection onto the route is denoted as the origin O. An extreme counter-example is that if the robot is running 100 meters per second and the target is transmitting 1 packet per second, we cannot reconstruct the target power space distribution $s(t)$ along the route because there are too few RSS samples. In reality, it is highly possible that the packet transmission rate of a target (e.g. a laptop) may be quite slow. Hence, a strict control of the moving velocity is necessary.

In practice, $s(t)$ can be much more complicated because of the multipath effect in an indoor environment. The theories presented in this paper actually give upper bounds of the robot velocity for reconstructing $s(t)$. Our experimental results in Section 4 match the theory very well.

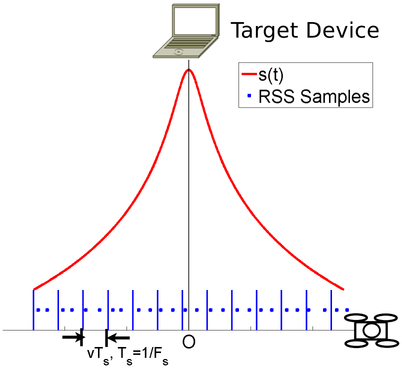


Fig. 1. Power Distribution s(t) Over a Route

3.3 Our Solution

We now present a fundamental theorem for a moving sniffer to correctly collect RSS samples, with which we can reconstruct the target's power distribution. For clarity and conciseness, we assume that a robot carries the sniffer and such a system produces a moving sniffer.

Recall Formula (1) gives the physical model of wireless signal attenuation. We define $s(t)$ as the power distribution over a route. In the discussion of RSS sampling, we ignore the noise term $X\sigma$ in Formula (1). This does not affect the essence of our sampling theory. Furthermore, noise is of high frequency and the sampling process filters a part of the noise.

We use the signal sampling theory [1] to address the problem. Considering the velocity of the robot and the target device packet transmission rate, we derive a RSS sampling theorem in Theorem 1.

Theorem 1. *A robot armed with a wireless sniffer moves at a velocity of v m/s along a route. The target wireless device transmits at a rate of F_{Pkt} packets per second. The collected RSS samples can be used to reconstruct the target power distribution along the route if and only if the following condition is satisfied,*

$$v < \frac{F_{Pkt} v_0}{2F_0}, \tag{2}$$

where v_0 is a baseline velocity, and F_0 is the bandwidth of $s(t)$ at velocity v_0, which is denoted as $s_0(t)$.

We use Figure 1 to explain how we prove Theorem 1. Recall $s(t)$ is the target power distribution $s(t)$ over the route, dots below $s(t)$ are RSS samples, and the origin O is the target's orthogonal projection on the route. We refer to the Nyquist sampling theorem and claim that, in order to reconstruct $s(t)$, the RSS sampling frequency F_s should be higher than twice of $s(t)$'s band limit F_{max}. Based on this point, it is obvious that the sniffer should sample in every space interval of length vT_s along the route, where $T_s = \frac{1}{F_s}$ is the sampling interval. In Figure 1, we place a set of short bars as boundaries of the space intervals along the route. The sniffer should collect at least one packet within each interval. Therefore, the robot should control its velocity and stay within each cell longer than the packet transmission interval. Since $s(t)$ scales over speed v, so does F_{max} vary over v. In order to represent F_{max} using speed v, we use Fourier transform's properties, select v_0 as a baseline velocity, and build the relationship among F_0, v_0, F_{max} and v, where F_0 is the bandlimit of $s(t)$ at speed v_0. Finally, we have Formula (2). The proof of Theorem 1 can be found in Appendix A of our technical report which is available on requirement.

Theorem 1 tells us that in order to balance accuracy and efficiency, we should control our robot to move reasonably fast, but below a velocity of $\frac{v_0 F_{Pkt}}{2F_0}$ m/s.

4 Evaluation

We conducted extensive experiments and simulations to verify the correctness of our RSS sampling theory. In this section, we first introduce our system BotLoc, a moving robot armed with a wireless sniffer for localization. Then, we present experimental results to validate our RSS sampling theory.

4.1 BotLoc System

Figure 2 shows the architecture of BotLoc while Figure 3 shows a photo of the BotLoc prototype. BotLoc consists of two main components: (i) robot subsystem, and (ii) positioning subsystem.

The robot subsystem drives the robot along a specified route by accepting velocity and heading commands from a software controller on a laptop, which steers the robot. The robot has an odometry coordinate system and returns its odometric pose to the controller program. The pose is represented as (x, y, θ), where x and y are the two-dimensional coordinates, and θ is the orientation of the robot with respect to its starting pose.

The positioning subsystem utilizes an antenna to sniff wireless packets and derive RSS samples. The localization analyzer correlates the collected RSS to the robot's location over time. Therefore, when the surveillance is finished, the localization analyzer inspects the signal strength time series coupled with position information in order to compute the target's location. Once the computation is finished, the positioning subsystem sends the result to the robot subsystem via the software controller, which drives the robot to the derived position. During the surveillance operation, the positioning system may also command the robot to adjust its speed and heading and cater to the target's transmission pattern. We now introduce BotLoc's configuration. The robot we choose is *Pioneer P3-DX* [16], a remotely operated and programmable robot. P3-DX can be controlled by either a joystick or a program. The robot is assembled with a laser rangefinder (*LMS200* [15]), a sonar array, and installed with a laser mapping and navigation system, which enables the robot to map the environment such as a building floor. The robot controller runs on a Lenovo W500 laptop, and it uses an object-oriented, robot control applications-programming interface (ARIA) from the manufacturer to control the robot. A robot operator can

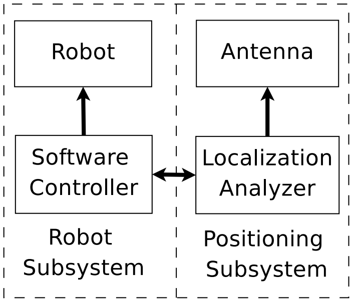


Fig. 2. *BotLoc* Function Modules

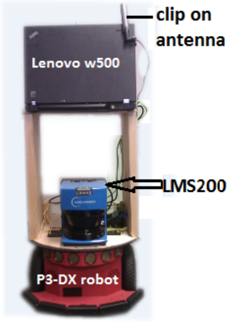


Fig. 3. Prototype

launch a client program such as *MobileEyes* [14] to set up a TCP connection to the robot controller, and execute localization tasks remotely.

The positioning subsystem uses a Ubiquiti 802.11a/b/g PCMCIA Cardbus connected to a clip-on antenna and runs a wireless sniffing program on the laptop. If we need to generate a map with AP positions marked, Kismet can collect wireless traffic of the whole spectrum via channel hopping.

4.2 Experiment Setup

We design real world experiments in two different environments: office, and track field.

In the *office case*, we conduct indoor localization. We place a laptop in one office room as a target, which keeps sending out ICMP packets at a constant rate of 100 packet/s. We drive BotLoc along the corridor along a straight line, and choose the position where the peak value of the target's power distribution is sensed to be the target's estimated position. Although this estimated position is not accurate enough to point out where the target stays in the room, this result is sufficient for us to illustrate the accuracy of our localization scheme via applying the RSS sampling theory. To determine the bound of robot's velocity, we calculate the target's average packet transmission rate, which is 111 packet/s. This rate is higher than 100 packet/s that we set to the target, because in the monitor mode, the sniffer is able to capture the re-transmitted frames.

To demonstrate how the robot's velocity affects localization accuracy, we synthesize multiple robot pose time series under different robot speeds, and calculate the estimated projection of targets using the synthesized data. With a group of robot pose time series collected from real-world experiments, our simulation follows four steps: (i) We select a group of velocities. (ii) For each velocity, we synthesize new robot pose time series from our real-world data, respectively. (iii) With synthesized pose time series and RSS time series, we use linear interpolation to identify the location with the maximal RSS readings. (iv) We determine the estimated projection point X'. We set velocity v to be 0.2, 2, 20, 40.0, 80.0, 160.0, 320.0, 640.0, 1280.0 , and 2560.0m/s, and run simulations for each value. The simulation results are presented in Figure 4.

In the case of *track field*, we conducted outdoor localization via the triangulation principle to actually derive the target's position. First, we choose a Nexus One smartphone as our target. We set the smartphone as an AP so that it keeps broadcasting beacon frames. Next, we design two perpendicular lines as our routes. We place the smartphone 1 meter away from both routes, and locate it via BotLoc. Then we draw two perpendicular lines to these routes with each line passing through the position where the maximum RSS is sensed along a route. The intersection of these two perpendicular lines is chosen as the target's position X'. Therefore, the localization error is the distance from this intersection X' to the target's real location. We adjust the distance between the smartphone and each of the routes to be 1, 2, 4, 8, 12 meters, and test 20 times for each distance respectively.

4.3 Performance of BotLoc

In the *office case*, we present our localization experiment results in Figure 4. Figure 4 shows that (i) when the velocity is below 320m/s, the average distance error is below two meters. This cutoff velocity is large because the target is transmitting very fast at 111 packets/s. It also explains why related work [11], [24] can derive seemingly correct results although they don't have the support of our RSS sampling theorem; (ii) The localization error changes with the robot velocity. These observations validate our RSS sampling theorem.

In the *track field case*, we present our localization experiment results via the triangulation strategy in Figure 5. The x-axis is the distance between the target and each route, while the y-axis is the localization error under each setting. The solid line represents the mean of localization error and the short segments show the confidence intervals with respect to every mean. As the distance increases, the confidence interval grows, too. When the distance is 12 meters, the mean is 2.4 meters. This is acceptable as it is easy for human beings to spot a mobile device within a range of 2.4 meters nearby.

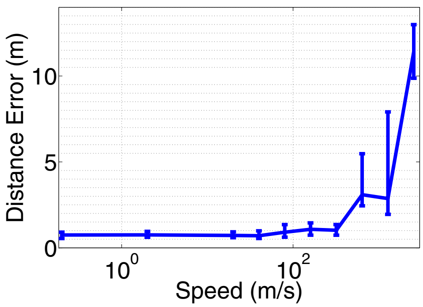


Fig. 4. Localization Error vs. Velocity

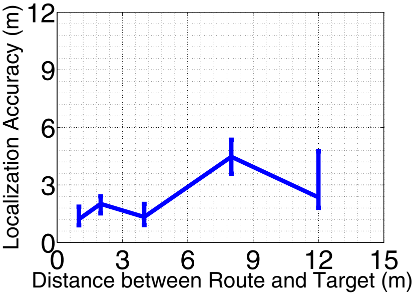


Fig. 5. Localization Accuracy with Triangulation

5 Related Work

There has been extensive work on device positioning in WiFi and sensor networks. Due to space limitations, we only review the existing work most related to our paper: Most existing techniques provide localization in a two-dimensional space (e.g., longtitude/latitude). Positioning systems are classified as outdoor and indoor systems, respectively, which feature vastly different requirements and techniques. The most popular outdoor positioning system is GPS [7]. Many cellular mobile networks also belong to this category, and allow the tracking of powered-on devices through the operator's base-transceiver stations. Indoor positioning systems include RADAR [3], LANDMARC [20], the digital Marauder's Map [10], Lighthouse [25], and VORBA [21]. All these systems position a mobile device based on the measured signal strength. In particular, the former three utilize a dense grid of omnidirectional base stations, while the latter two rely on base stations with revolving unidirectional antennas. Active Badge [31], Active Bat [13] and Cricket [23] can provide better localization accuracy than outdoor systems due to the usage of a large number of positioning-support sensors.

Subramanian *et al.* [27] use electronically steerable Phocus Array antennas from Fidelity Comtech [9] for wardriving and collecting RSSs with directional information. Han *et al.* [12] take RSS samples from wardriving and use the gradient information derived from RSS to infer the position of an AP.

The indoor localization technique using smartphones is widely studied by researchers. Point Inside [17] and Google Maps for Android [19] are smartphone applications developed for malls, retailers, and airports. Point Inside helps navigate users to stores or products of interest. Google Maps for Android shows a user his or her location on an indoor map. [29] utilizes acoustic background spectrum (ABS) as a room fingerprint to determine which room a smartphone holder visits. Quigley *et al.* [24] leverage robots to generate indoor maps with off-the-shelf SLAM techniques and use multiple sensors to collect training data including WiFi signals, camera images, and locate a human by searching for closest matched training data. Constandache *et al.* [5] creatively adopt audio beacons to navigate people via mobile phones. They capture users' movement traces using accelerometer and compass measurements and establish a global view of users' positions in a remote server. This server provides routes to users to reach the people of interest.

6 Conclusion

In this paper, we have addressed a critical, but unanswered question in a forensic wireless localization scene: How can we use collected RSS samples by a moving sniffer, carried by humans or vehicles, to reconstruct the target transmitter's power distribution in the space? With a reconstructed RSS distribution, we can derive the location yielding the maximum RSS. Our experiments show that in a complicated environment such as building corridors, triangulation cannot be used for accurate localization via single device. A simple approach, which is better and more robust, is to use where the maximum RSS is sensed as the target's location.

We innovatively apply the Nyquist sampling theory to address this problem and develop our RSS sampling theory. We demonstrate the use of this RSS sampling theory via precise localization using a moving robot. We developed a fully functional localization systems: BotLoc. BotLoc is a P3-DX robot armed with a wireless sniffer. It applies our RSS sampling theory for collecting RSS time series and locating a target mobile device with a moving robot armed with a wireless sniffer. It is infrastructure-free and training-free. Our contribution also includes extensive experimental results which demonstrate the effectiveness of the proposed theorem and systems. We expect that our theory and system provide a fundamental empirical and theoretical framework for forensic wireless localization via moving sniffers.

References

[1] Oppenheim, J.R.B.A.V., Schafer, R.W.: Discrete-time signal processing. Prentice Hall (1999)

[2] Anderson, N.: Swat team throws flashbangs, raids wrong home due to open wifi network (2012), `http://arstechnica.com/tech-policy/2012/06/swat-team-throws-flashbangs-raids-wrong-home-due-to-open-wifi-network/`

[3] Bahl, P., Padmanabhan, V.N.: RADAR: An in-building RF-based user location and tracking system. In: Proceedings of INFOCOM (2000)

[4] Chan, C.: The swat team gently reminds a girl to secure her wi-fi network by raiding her house with flashbangs (2012), `http://gizmodo.com/5922322/the-swat-team-gently-reminds-a-girl-to-secure-her-wi+fi-network-by-raiding-her-house-with-flashbangs`

[5] Constandache, I., Bao, X., Azizyan, M., Choudhury, R.R.: Did you see bob?: human localization using mobile phones. In: Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking (2010)

[6] Durgin, G.D., Rappaport, T.S., Xu, H.: Measurements and models for radio path loss and penetration loss in and around homes and trees at 5.85 ghz. ACM Transactions on Communications 46(11), 1484–1496 (1998)

[7] Enge, P., Misra, P.: Special issue on global positioning system. Proceedings of the IEEE 87(1), 3–15 (1999)

[8] Faria, D.B.: Modeling signal attenuation in ieee 802.11 wireless lans - vol. 1. In: Stanford University, Tech. Rep. (July 2005)

[9] Fidelity Comtech, Inc. 802.11 phocus array antenna system by fidelity comtech (2009), `http://www.fidelity-comtech.com/`

[10] Fu, X., Zhang, N., Pingley, A., Yu, W., Wang, J., Zhao, W.: The digital marauders map: A new threat to location privacy in wireless networks. In: Proceedings of ICDCS (2009)

[11] Haeberlen, A., Rudys, A., Flannery, E., Wallach, D.S., Ladd, A.M., Kavraki, L.E.: Practical robust localization over large-scale 802.11 wireless networks. In: Proceedings of the 10th Annual International Conference on Mobile Computing and Networking MOBICOM (2004)

[12] Han, D., Andersen, D.G., Kaminsky, M., Papagiannaki, K., Seshan, S.: Access point localization using local signal strength gradient. In: Moon, S.B., Teixeira, R., Uhlig, S. (eds.) PAM 2009. LNCS, vol. 5448, pp. 99–108. Springer, Heidelberg (2009)

[13] Harter, A., Hopper, A., Steggles, P., Ward, A., Webster, P.: The anatomy of a context-aware application. In: Proceedings of MOBICOM (1999)

[14] A. T. Inc. Mobileeyes (2011), `http://www.mobilerobots.com/ResearchRobots/PioneerSDK/MobileEyes.aspx`

[15] A. T. Inc. Laser navigation package (2012), `http://www.mobilerobots.com/ResearchRobots/Accessories/LaserNavigationPkg.aspx`

[16] A. T. Inc. Pioneer p3-dx (2012), `http://www.mobilerobots.com/researchrobots/researchrobots/pioneerp3dx.aspx`

[17] P. I. Inc. Pointinside (2012), `http://www.pointinside.com/`

[18] KTRK-TV/DT. Man accused of downloading child porn over his neighbor's wi-fi (2012), `http://abclocal.go.com/wabc/story?section=news/local&id=8917541`

[19] McClendon, B.: A new frontier for google maps: mapping the indoors (2012), `http://googleblog.blogspot.com/2011/11/new-frontier-for-google-maps-mapping.html`

[20] Ni, L.M., Yiu, Y.L., Lau, C., Patil, A.P.: LANDMARC: Indoor location sensing using active RFID. In: Proceedings of PerCom, pp. 407–415 (2003)

[21] Niculescu, D., Nath, B.: VOR base stations for indoor 802.11 positioning. In: Proceedings of MOBICOM (2004)

[22] Poulsen, K.: Record 13-year sentence for hacker max vision (2010), `http://www.wired.com/threatlevel/2010/02/max-vision-sentencing/`

[23] Priyantha, N.B., Chakraborty, A., Balakrishnan, H.: The Cricket Location-Support System. In: Proceedings of MOBICOM (2000)

[24] Quigley, M., Stavens, D., Coates, A., Thrun, S.: Sub-meter indoor localization in unmodified environments with inexpensive sensors. In: IROS, pp. 2039–2046. IEEE (2010)

[25] Römer, K.: The lighthouse location system for smart dust. In: Proceedings of MobiSys (2003)

[26] Jones, S.E.: Internet access stealing becoming a major law enforcement concern (2012), `http://voices.yahoo.com/internet-access-stealing-becoming-major-law-enforcement-11216494`

[27] Subramanian, A.P., Deshpande, P., Gao, J., Das, S.R.: Drive-by localization of roadside wifi networks. In: Proceedings of INFOCOM (2008)

[28] Sylvain, A.: Internet thieves piggyback on legitimate users (2012), `http://www.usatoday.com/tech/news/story/2012-04-08/internet-theft-web/54116488/1`

[29] Tarzia, S.P., Dinda, P.A., Dick, R.P., Memik, G.: Indoor localization without infrastructure using the acoustic background spectrum. In: Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services (2011)

[30] Thompson, C.: Innocent man accused of child pornography after neighbor pirates his wifi (2011), `http://www.huffingtonpost.com/2011/04/24/unsecured-wifi-child-pornography-innocent_n_852996.html`

[31] Want, R., Hopper, A., Falcao, V., Gibbons, J.: The active badge location system. ACM Transactions on Information Systems 10(1) (January 1992)

[32] Yang, S., Kurose, J., Levine, B.N.: Disambiguation of residential wired and wireless access in a forensic setting. In: Proceedings of INFOCOM (2013)

An Optimal Leakage Detection Strategy for Underground Pipelines Using Magnetic Induction-Based Sensor Networks

Xin Tan and Zhi Sun

Department of Electrical Engineering, State University of New York at Buffalo, U.S.
{xtan3,zhisun}@buffalo.edu

Abstract. It is difficult to detect small leakages in underground pipelines with high accuracy and low-energy cost due to the inaccessible underground environments. To this end, the Magnetic Induction (MI)-based wireless sensor network for underground pipeline monitoring (MISE-PIPE) is introduced in [13]. The MISE-PIPE deploys high-density underground MI sensors along the pipelines, which provide necessary measurements for leakage detection with very high resolution. However, in order to provide accurate and low-cost leakage detection based on MISE-PIPE, an optimal deployment and activation strategy for those sensors is needed. In this paper, we provide an optimal strategy to detect leakages in underground pipelines based on the MISE-PIPE framework. Based on the proposed detection strategy, the error bound is derived to characterize the accuracy, while the energy consumption is analyzed to model the system energy cost. By trading off the accuracy and the energy consumption, an optimization function is developed to achieve the optimal performance.

Keywords: Underground pipelines, leakage detection and localization, deployment and activation of sensors, estimation accuracy, energy consumption.

1 Introduction

Leaks of hazardous liquid such as crude oil, gasoline can cause serious environmental pollution if the leakages are not quickly detected and repaired. Even though the underground pipeline networks are the safest way to transport fluids for long distances, at least one obvious leakage occurs on a long underground pipeline every year [13]. As introduced in [1], [3], [4], [5], etc., different types of technique for leakage detection are developed. However, existing methods cannot provide accurate measurements with high enough resolution to detect small leakages due to the inaccessible underground environments. Moreover, only a single type of the sensing technique is utilized by existing leakage detection systems, which deteriorates either the detection accuracy or the energy efficiency.

In [13], a new framework called Magnetic Induction (MI)-based wireless sensor network for underground pipeline monitoring (MISE-PIPE) is introduced, where underground sensor nodes are densely deployed along the underground pipelines to provide high resolution measurements. The sensing data are transmitted through the

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 414–425, 2013.

underground wireless sensor nodes using the Magnetic Induction (MI)-based communication mechanism. Moreover, multiple detection and localization techniques can be organized and concerted together so that those techniques can make a complement to each other. The MISE-PIPE system detects and locates the leakage by jointly using two types of sensor: The sensors for rough detection with low energy consumption; the sensors for accurate detection with high-energy consumption. The imprecise but low-cost technique is used to keep monitoring the status of pipelines and activate the accurate detection with high-energy consumption if further information is needed.

Although the MISE-PIPE concept has promising potentials, it is far from mature and several key challenges need to be solved: First, since more than one type of sensors are adopted in the system, a suitable coordination strategy needs to be designed to optimally utilize the multiple detection and localization techniques. Second, rigorous mathematical models for detection accuracy and system energy consumption need to be derived to prove the optimality of the proposed strategy. In this paper, we propose an optimal strategy to detect leakages in underground pipelines based on the MISE-PIPE framework, which can address the above challenges.

The remainder of this paper is organized as follows: The existing pipeline leakage detection techniques are summarized in section 2. The system architecture and localization methods we use for this system are presented in section 3. The optimal strategy, including estimation accuracy and energy consumption, is introduced in section 4 and we make a conclusion in section 5.

2 Existing Pipeline Monitoring Techniques

Transient-Based Methods: Transient-based methods have been intensively analyzed by the research community recently. It has been analyzed in [1], [3], [5], [6] and [8]. Transient-based methods are usually based on deploying pressure sensors at checkpoints to measure the pressure change triggered by transient events in pipelines such as opening/closing a valve or starting-up or shutting down a pump. Therefore, the size and location of the leakage can be identified by transient-based methods. However, this technique cannot provide enough leakage detection accuracy since the pressure sensors can be only deployed in limited number of locations such as checkpoints.

Acoustic Methods: Acoustic sensors can be used to trace the vibration data to its source to detect the small leakage along pipelines. This technique is introduced in [4]. However, in order to cover the whole pipeline network, a high density of acoustic sensors need to be deployed along the pipelines due to the limitation of the detection range. It is impossible for underground pipeline network due to the deployment and maintenance difficulties. Moreover, for large leaks that will not generate vibrations in the characteristic high frequencies, acoustic sensors are not sensitive. Thus, the acoustic sensors can be only installed in limited locations such as checkpoints to accurately detect the small leakages near them.

Soil Properties Methods: The leakage on pipelines can be detected through the identification of abnormal value of the soil properties since the fluid leaked from the pipelines will change the soil properties around the leakage. For example, hydrocarbon vapor sensors can be used to detect the leakage on pipeline carrying liquefied natural gas [10]. Soil dielectric property sensors can be used to detect the leakage of pipeline transporting crude oil [14]. Temperature sensors are usually used to detect hot liquid leaks that can change the temperature of surroundings [9]. The techniques based on soil property sensors can provide accurate leakage detection and localization. However, currently the wire-based communication system is used to transmit the measurements derived by the underground soil property sensors to remote administration center [14]. Moreover, to cover the whole pipeline networks, we need to deploy soil property sensors along the whole networks to monitor every interval of the pipelines. Thus, the deployment cost of the wire-based system is extremely high and the energy we need to keep those sensors working is also unaffordable.

3 System Architecture and Localization Methods

MISE-PIPE has a clustered architecture of heterogeneous sensors, which consists of two layers: the hub layer and the in-soil sensor layer [13]. As shown in Fig.1, the hub layer consists of the sensors for rough detection and estimation that are deployed inside the pipeline at the checkpoint or the pump stations. The in-soil sensor layer consists of the sensors for accurate localization that are deployed along the underground pipelines. The pipelines are divided by the checkpoints or pump stations [13].

Sensors in hub layer for rough detection and estimation in MISE-PIPE system are used to find out suspicious leak points so that they are not required to have precise measurements. The primary requirement should be low energy cost. By considering these requirements comprehensively, pressure sensors and acoustic sensors can be deployed in hub layer. The pressure sensors utilize the transient-based method to identify the area where the pipelines are likely to have leakages. Acoustic sensors can be utilized as a complement to the pressure sensors since they can have an accurate detection and localization but very small detection range.

The working phases of MISE-PIPE are introduced in [13]. In normal case, pressure and acoustic sensors in the hub layer keep working to collect data along the pipeline. Sensors at checkpoints and pump stations are connected to an aboveground gateway and the data achieved by those sensors can be transmitted to the remote administration.

Once an alarm is achieved by the remote administration center, the sensors in hub layer of the pipelines will be notified by the administration center. The sensors in hub layer send out data requests to the sensors in in-soil layer. Those sensors equipped with MI coils can be woken up by the request from the hubs and collect measurements along the pipeline. The measurements are sent to the sensors in hub layer. After processing these measurements in the hubs, the accurate location of the leakage can be determined and the hubs will transmit the result to the administration center [13].

Temperature sensors can be deployed along the pipeline for accurate detection and localization. In a certain depth of ground, the temperature is not affected by climate change a lot and the year-end change is less than 1 centigrade. The underground temperature can be only affected by the fluids in the underground pipelines. The temperature distribution is steady if the pipeline is intact, but it will change if there is a leak. As shown in Fig.1, temperature sensors with MI transceiver are attached on the outside surface of pipeline. If the pipe is intact, as shown in Fig.2, the heat is uniform along the pipeline and the temperature follows a uniform distribution (Curve 1). If there is a leak, the fluid will flow out and change the temperature of the outside surface of pipe wall (Curve 2). Heat from the leak point can spread along the pipeline fast so that the temperature sensors can get measurements. Temperature sensors are not used to detect the temperature of soil but the temperature change of outside surface of pipe wall. That is because the material of pipe wall has a higher thermal conductivity so that the heat can spread faster and more effective measurements can be achieved by temperature sensors.

The leak on pipeline can be regarded as a constant heat source and the temperature change along the pipeline follows the Gaussian distribution. According to [7], the temperature change along the pipeline can be written as a function shown as follows:

$$\Delta T(x) = T_1(x) - T_0(x) = ae^{-\frac{(x-b)^2}{c}}\ , \tag{1}$$

where x is the position on pipeline, $\Delta T(x)$ is the temperature change at point x and it can be represent as the difference between final status temperature $T_1(x)$ and initial temperature $T_0(x)$. a, b and c are the parameters for the Gaussian distribution that need estimated. Obviously, parameter b tells us the location of the leakage.

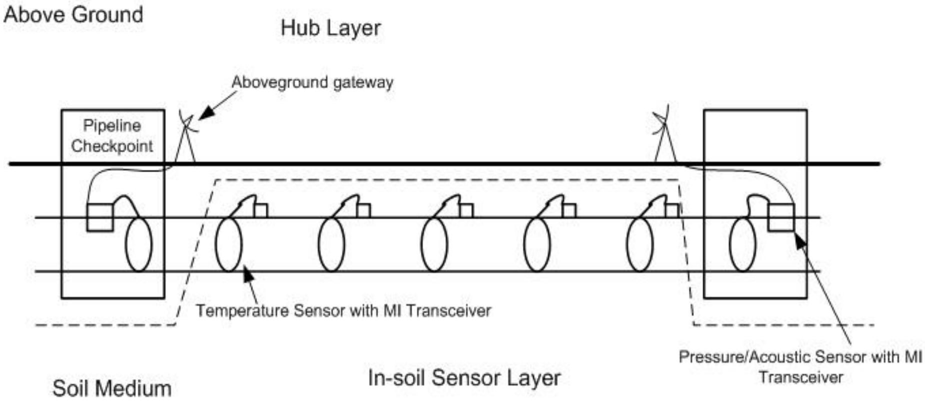


Fig. 1. System architecture of MISE-PIPE

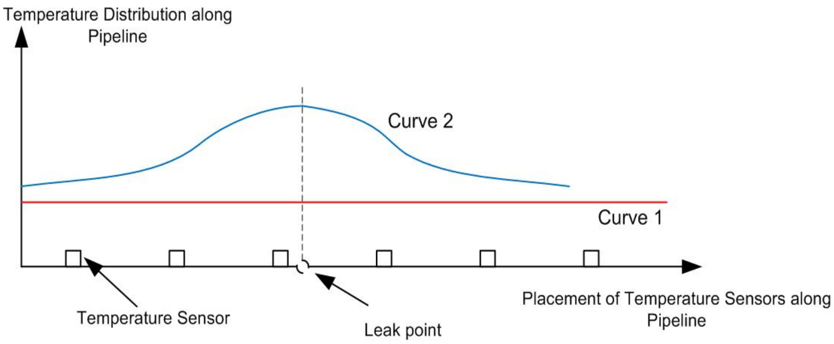


Fig. 2. Temperature Distribution along Pipeline

4 Optimal Strategy for Deployment and Activation of Sensors

4.1 Communication Range

Temperature sensors in in-soil layer are all equipped with MI transceivers. Hence, the deployment of sensors depends on the communication range of those transceivers. In fact, sensors equipped with MI transceivers cannot be very far away from each other because of the path loss and this becomes an restricted condition of the our strategy. Assuming that the distance between two neighbor sensors is r, the angle frequency of the transmitting signal is ω. According to [11] and [12], the path loss of the MI waveguide can be expressed as

$$L_{MI}(r,\omega) \cong 6.02 + 20lg\frac{Z}{\omega M} , \tag{2}$$

where M is the mutual induction between the adjacent sensors. As discussed in [12], M can be easily deduced by the magnetic potential of the magnetic dipole

$$M \cong \frac{\mu\pi N^2 a^4}{4r^2} , \tag{3}$$

where μ represents the permeability of the pipeline, N is the turns of wire on MI coils and a is the coil radius.

According to the result in [11], to guarantee the transmitting signal received by the nearest receiver, the path loss should meet the following inequation.

$$P_t - L_{MI}(r, \omega_0 + 0.5B) \geq P_{th} , \tag{4}$$

where P_t is the transmission power, P_{th} is the minimum power for correct demodulate a signal, ω_0 is the central frequency and B is the required bandwidth between two sensors. Thus, by combining (2) (3) (4), the distance between to neighbor sensors r is bounded by

$$r \leq (\frac{\mu\pi N^2 a^4 \omega}{4Z})^{\frac{1}{3}} 10^{\frac{P_t - P_{th} - 6.02}{60}} . \tag{5}$$

According to (5), if the material and structure of pipelines and MI transceivers are fixed, we can get the maximum distance between two neighbor sensors. The strategy for deployment and activation of sensors is bounded by this maximum value. By using the parameters of the MI waveguide developed in [12], we can numerically calculate the maximum distance $r \leq 92$ meters.

In fact, sometimes we don't need to activate sensors with such a high density if the leakage is very large. As shown in Fig.3a, we can activate every other, one in every three, or even on in every i sensors. Thus, those nodes between two activated sensors become the relay coils that will not generate any power or get any measurements but just help the signal transmission. Energy will be saved if we do so for large leakage.

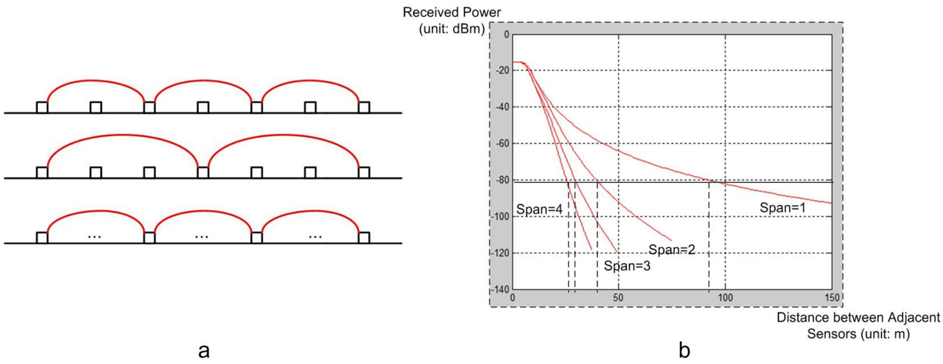


Fig. 3. Activation of sensors for large leakage

In this case, the bound of the distance between two neighbor sensors will be changed since the distance between two transceivers will be $d = ir$, where r is the distance between two neighbor sensors and i is the span between two transceivers (the span is i if we start one in every i sensors and $i-1$ nodes become relay coils). According to [11], the path loss with a span of i can be expressed as

$$L_{MI}(r,i,\omega) \cong 6.02 + 20lg\zeta(\frac{Z}{\omega M},i) \ , \tag{6}$$

where the polynomial $\zeta(x,i)$ can be developed as

$$\zeta(x,1) = x \ , \ \zeta(x,2) = x^2 + 1 \ ,\ldots, \ \zeta(x,i) = x\zeta(x,i-1) + \zeta(x,i-2) \ . \tag{7}$$

Thus, the distance between two sensors with a span of i will be bounded by

$$P_t - L_{MI}(r,i,\omega_0 + 0.5B) \geq P_{th} \ . \tag{8}$$

By using the same parameters developed in [11], we set the minimum received power -80 dBm. The bound of distance r can be determined for different value of i. For different span i, the maximum distance between two neighbor sensors can be achieved in Fig.3b. From the figure we can get that with the span increasing, the maximum distance between two sensors will be smaller.

4.2 Estimation Accuracy

The estimation accuracy becomes an important factor of optimal strategy. We can use the discrete temperature signals achieved by temperature sensors to fit a Gaussian function and estimate the mean value of its independent variable. As shown in Fig.4, the original point O is the suspicious leak point located by rough detection based the sensors in checkpoints or pump stations. Obviously, the actual location of leak point has a deviation from the original point O. Assume that m sensors are activated on both sides of point O, s_i represents the position of the i th activated sensor on the right of point O, s_{-i} represents the position of the ith activated sensor on the left of point O.

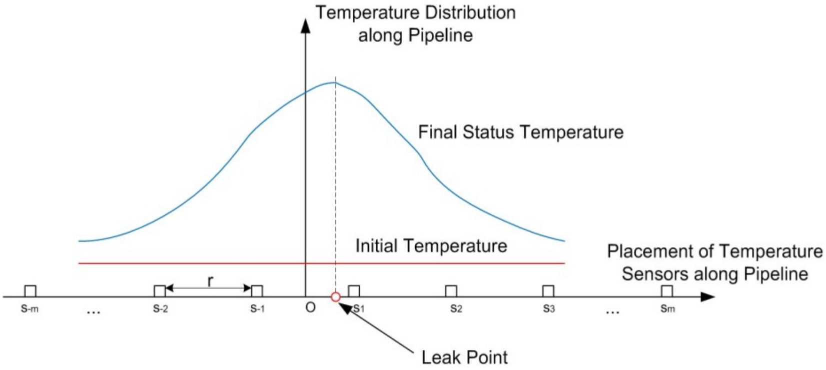

Temperature Distribution along Pipeline
Final Status Temperature
Initial Temperature
Placement of Temperature Sensors along Pipeline
r
S-m
...
S-2
S-1
O
S1
S2
S3
...
Sm
Leak Point

Fig. 4. The placement of temperature sensors for temperature detection

As discussed in section 2, the temperature change along pipeline follows a Gaussian distribution and it can be represented as the difference between final statues temperature and initial temperature as (1). If the distance between two neighbor activated sensors is d, the position of each activated temperature sensor can be rewritten as

$$s_n = s_1 + nd \qquad n = -m, -(m-1), \dots, -1, 0, 1, \dots, m-1\ . \tag{9}$$

Notice that here d represents the distance between activated sensors. That means, there may be some other inactivated nodes among them to work as relay coils without sensing or energy supply. Thus, (1) can be rewritten as

$$\Delta T(s_1+nd) = T_1(s_1+nd) - T_0(s_1+nd) = ae^{-\frac{(s_1+nd-b)^2}{c}}\ . \tag{10}$$

We set $\hat{b}$ as the estimation of the leak point position b. Then the minimum variance of $\hat{b}$ can be determined by calculating the Cramer-Rao lower bound.

Consider the multiple observations

$$\Delta T[n] = T_1[n] - T_0[n] = ae^{-\frac{(s_1+nd-b)^2}{c}} + w[n]$$
$$n = -m, -(m-1), \dots, -1, 0, 1, \dots, m-1\ . \tag{11}$$

where $w[n]$ is the white Gaussian noise with variance σ^2 and we have $\sigma^2 = \sigma_1^2 + \sigma_0^2$. Here, σ_1^2, σ_0^2 are the variance of noise added on $T_1[n]$ and $T_0[n]$. According to (11), the likelihood function becomes

$$P(\Delta T;\vec{\theta}) = \frac{1}{(2\pi\sigma^2)^m}\exp\{-\frac{1}{2\sigma^2}\sum_{n=-m}^{m-1}(T[n] - ae^{-\frac{(s_1+nd-b)^2}{c}})^2\}\ , \tag{12}$$

where $\vec{\theta} = [a \quad b \quad c]^T$ is the parameter vector need estimated. The log-likelihood function can be achieved by taking logarithm on both sides of the equation

$$lnP(\Delta T;\vec{\theta}) = ln\frac{1}{(2\pi\sigma^2)^m} + ln\{-\frac{1}{2\sigma^2}\sum_{n=-m}^{m-1}(T[n] - ae^{-\frac{(s_1+nd-b)^2}{c}})^2\}\ . \tag{13}$$

Since we have three elements in the parameter vector, a three-order Fisher information matrix need to be calculated to determine the error bound. Assume that the Fisher information matrix can be written as

$$I(\vec{\theta}) = \begin{pmatrix} x_{aa} & x_{ab} & x_{ac} \\ x_{ba} & x_{bb} & x_{bc} \\ x_{ca} & x_{cb} & x_{cc} \end{pmatrix} , \qquad (14)$$

where

$$x_{ij} = -E[\frac{\partial^2 lnP(\Delta T;\vec{\theta})}{\partial i \partial j}] \qquad\qquad i,j \in \{a,b,c\} \ . \qquad (15)$$

By calculating the derivatives we can find that if we activate roughly equal numbers of sensors on both sides of the leak point, we have

$$x_{ab} = x_{ba} \cong 0 \ , \ x_{bc} = x_{cb} \cong 0 \ . \qquad (16)$$

Thus the Fisher information matrix can be written as

$$I(\vec{\theta}) = \begin{pmatrix} x_{aa} & 0 & x_{ac} \\ 0 & x_{bb} & 0 \\ x_{ca} & 0 & x_{cc} \end{pmatrix} , \qquad (17)$$

where $x_{ac} = x_{ca}$. Inverting the matrix yields

$$I(\vec{\theta}) = \begin{pmatrix} \frac{x_{cc}}{x_{aa}x_{cc}-x_{ac}^2} & 0 & -\frac{x_{ac}}{x_{aa}x_{cc}-x_{ac}^2} \\ 0 & \frac{1}{x_{bb}} & 0 \\ -\frac{x_{ac}}{x_{aa}x_{cc}-x_{ac}^2} & 0 & \frac{x_{aa}}{x_{aa}x_{cc}-x_{ac}^2} \end{pmatrix} . \qquad (18)$$

Thus, the Cramer-Rao lower bound for estimated leak point $\hat{b}$ is determined by

$$var(\hat{b}) \geq \frac{1}{x_{bb}} = \frac{1}{-E\left[\frac{\partial^2 lnP(\Delta T;\vec{\theta})}{\partial i \partial j}\right]} = \frac{c\sigma^2}{4a^2 \sum_{n=-m}^{m-1}(s_1+nd-b)^2 e^{-\frac{2(s_1+nd-b)^2}{c}}} \ . \qquad (19)$$

Once a leakage occurs, the parameters a, b, c are fixed. The variance of the noise σ^2 is also fixed. s_1 is the position of the first temperature sensor on the right of suspicious leak point. Since the suspicious leak point is located by rough detection, s_1 is fixed when accurate detection started. Thus, from the expression of error bound we can get that the Cramer-Rao lower bound for $\hat{b}$ depends on the number of sensors activated for detection $2m$ and the distance between two neighbor sensors d. Our target is to minimize this lower bound.

As shown in Fig.5, we set a temperature distribution function with central temperature change $a = 10\ ^oC$. For a pipeline with a length of 100 meters, we set the error of the rough localization $b = 10$ meter, the variance of the distribution $c = 1000$, the position of the first sensor $s_1 = 5$ m, the variance of the noise $\sigma = 2$. The figure shows the error bounds for different pairs m and d.

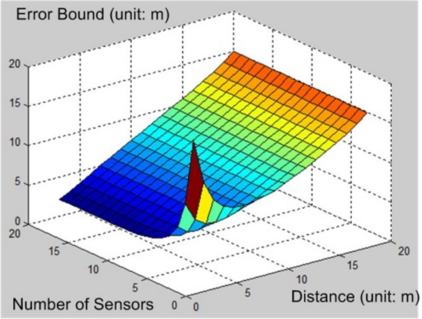


Fig. 5. Error bound varies with number of sensors and distance between two sensors change

From Fig.5 we can get that, with the number of sensors and the distance between activated sensors decreasing, we can get more accurate estimation since we deploy and activate a large number of sensors with a high density. Obviously, error bound can be minimized in this way but it is not an optimal strategy if we consider the energy consumption.

4.3 Energy Consumption

Since the sensors with MI transceivers are charged by batteries, the energy consumption becomes another aspect need to be considered. Assume that the energy cost of activating a temperature sensor with MI coils once is C, we start $2m$ sensors for accurate localization as discussed in section 4.2. The energy consumption becomes $2mC$. Notice that measurements are required to be sent to the nearest checkpoint or pump station which has a gateway above ground, we need to activate more nodes for communication. As shown in Fig.6, we choose to transmit the measurements to the left checkpoint since the suspicious leak point located by rough detection is close to the left checkpoint. Besides $2m$ nodes for accurate detection, a number of nodes on the left are activated to keep touch with the left checkpoint. In fact, we need to activate all the sensors between the suspicious leak point and the left checkpoint since we need to transmit the measurements based on them. Thus the total energy consumption becomes

$$C_{total} = (m + N)C \ . \tag{20}$$

Here we can assume $N \geq m$ since if $N < m$, which means there are less than m nodes between the suspicious leak point and the left checkpoint, the leak point is too close to the left checkpoint and it can be easily detected and accurately located by acoustic methods and there is no need to activate temperature sensors for further localization. Once the suspicious leak point is located by rough detection, the number of nodes on left N is fixed so that the total energy consumption is only related to m.

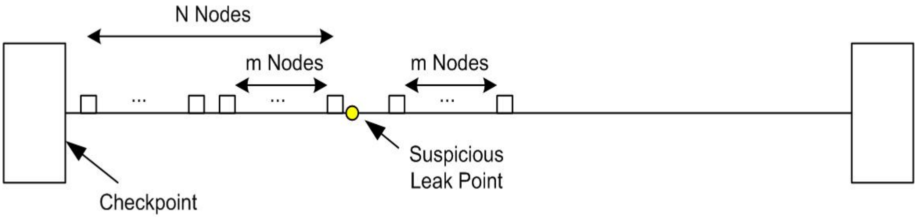


Fig. 6. Number of nodes activated for detection and communication

4.4 Optimal Strategy

To find out an optimal strategy for deployment and activation of sensors with MI transceivers, the accuracy and energy consumption becomes the two primary factors need balanced. As discussed in section 4.2, the estimation accuracy is a function of the distance between adjacent activated sensors d and the number of temperature sensors started for detection $2m$, which can be written as

$$f(d,m) = \frac{c\sigma^2}{4a^2 \sum_{n=-m}^{m-1}(s_1+nd-b)^2 e^{-\frac{2(s_1+nd-b)^2}{c}}} \ . \tag{21}$$

As discussed in section 4.3, the total energy consumption is a function of m that can be represented as

$$g(m) = (m+N)C \ . \tag{22}$$

By considering both two aspects, an optimization function can be given as

$$J = K_1 f(d,m) + K_2 f(m) \ , \tag{23}$$

where K_1, K_2 denote the weights of two parts. The restricted condition that depends on the communication range is discussed in section 4.1. Thus, the optimal strategy can be achieved by finding out the optimal solutions d^* and m^* to minimize the function

$$J^* = K_1 f(d^*,m^*) + K_2 g(m^*). \tag{24}$$

As shown in Fig.7a, we set $K_1 = 1$ and $K_2 = 1$. We utilize the same parameters discussed in section 4.2 and set the unit energy consumption $C = 1.5$ Joule. The minimum value J^* and optimal choices d^* and m^* can be achieved intuitively at the lowest point of the curved surface.

By calculating the values of J for different d and m by Matlab, we can get the minimum value $J^* = 31.6608$ if we activate totally 12 sensors with a distance of about 8 meters for detection.

Once a larger leakage is achieved by the rough detection and localization, the optimal choices of d and m will change. Fig.7b shows the diagram of optimization function for a larger leakage ($a = 20\ ^oC$,$C = 5000$). We can get the minimum value $J^* = 33.1965$ if we activate totally 12 sensors with a distance of about 13 meters.

For different magnitude of leakages, the optimal number of sensors that need activated can be determined in this way. Once the sensors with transceivers are deployed along the pipelines, we cannot change the distances between them. However, as discussed in section 4.1, we can find out the optimal span for sensor activation to fit the optimal distance d^*. For example, if we deploy the sensors with a distance of 7 meters, for the first small leakage, we can activate the sensors one by one since the optimal distance $d^* = 8$ meters. For the second case of large leakage, we can activate one in every two sensors so that the distance between two activated sensors will be $2 \times 7 = 14$ meters, which is close to the optimal distance $d^* = 13$ meters.

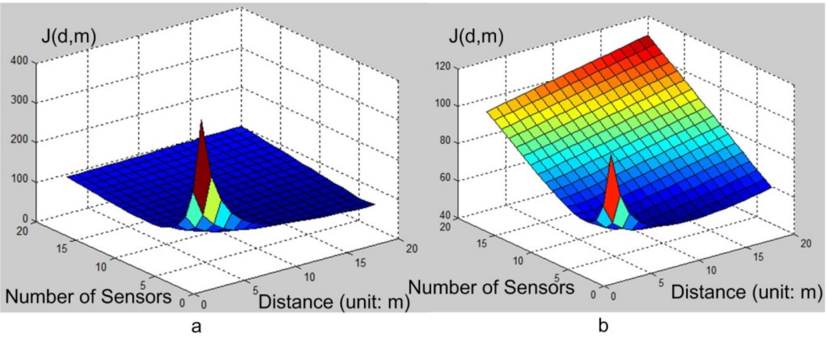


Fig. 7. Optimization function varies with number of nodes and distance between two nodes change

According to the data developed in [14], for some traditional detection and localization techniques, the relationship between leak size and localization accuracy is shown in Fig.8. For a small leakage that is less than 5% of the main flow, the localization error will be over 10%. That means, for a pipeline with a length of 100 meters, the error will be over 10 meters. If we use the traditional techniques for rough detection and localization which may cause an error of 10 meters, by using accurate localization with an optimal strategy, the error will decrease to 7.66 meters. The error can be reduced more if we increase the weight of accuracy K_1 in the optimization function.

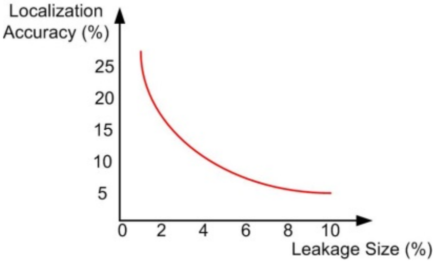


Fig. 8. Localization accuracy against leak size

5 Conclusion

In this paper, we propose an optimal leakage detection strategy for underground pipelines using the magnetic induction-based wireless network framework. We analyze the Cramer-Rao lower bound of estimated leak point, the communication range of MI transceivers, and the energy consumption for this sensor network. By considering these three aspects, an optimization function is developed and the optimal strategy can be achieved by minimize the function.

References

1. Ahadi, M., Bakhtiar, M.S.: Leak detection in water-filled plastic pipes through the application of tuned wavelet transforms to acoustic emission signals. Applied Acoustics 71, 634–639 (2010)
2. Beck, S., Curren, M., Sims, N., Stanway, R.: Pipeline Network Features and Leak Detection by Cross-Correlation Analysis of Reflected Waves. Journal of Hydraulic Engineering 131(8), 715–723
3. Colombo, A.F., Lee, P., Karney, B.W.: A selective literature review of transient-based leak detection methods. Journal of Hydro-environment Research 2(4) (April 2009)
4. Gao, Y., Brennan, M.J., Joesph, P.F., Muggleton, J.M., Hunaidi, O.: On the selection of acoustic/ vibration sensors for leak detection in plastic water pipes. Journal of Sound and Vibration 283(3-5) (May 20, 2005)
5. Giustolisi, O., Savic, D., Kapelan, Z.: Pressure-driven demand and leakage simulation for water distribution networks. Journal of Hydraulic Engineering 134(5), 625–635 (2008)
6. Guo, X.L., Yang, K.L., Li, F.T., Wang, T., Fu, H.: Analysis of first transient pressure oscillation for leak detection in a single pipeline. Journal of Hydrodynamics 24(3), 363–370 (2012)
7. Manca, O., Morrone, B., Naso, V.: Quasi-steady-state Three-dimensional Temperature Distribution Induced by a Moving Circular Gaussian Heat Source in a Finite Depth Solid. International Journal of Heat and Mass Transfer. 38(7) (May 1995)
8. Misiunas, D., Vitkovsky, J., Olsson, G., Simpson, A., Lambert, M.: Pipeline Break Detection Using Pressure Transient Monitoring. Journal of Water Resources Planning and Management 131(4) (July 2005)
9. Simon, I., Arndt, M.: Thermal and gas-sensing properties of a micromachined thermal conductivity sensor for the detection of hydrogen in automotive applications. Sensors and Actuators A: Physical 97-98 (April 1, 2002)
10. Sperl, J.L.: System pinpoints leaks on Point Arguello offshore line. Oil and Gas Journal 89(36)
11. Sun, Z., Akyildiz, I.F.: Deployment Algorithms for Wireless Underground Sensor Networks using Magnetic Induction. In: Global Telecommunications Conference, 2010. IEEE (2010)
12. Sun, Z., Akyildiz, I.F.: Magnetic Induction Communications for Wireless Underground Sensor Networks. IEEE Transactions on Antennas and Propagation 58(7) (July 2010)
13. Sun, Z., Wang, P., Vuran, M.C., AI-Rodhhan, M.A., AI-Dhelaan, A.M., Akyildiz, I.F.: MISE-PIPE: Magnetic induction-based wireless sensor networks for underground pipeline monitoring. Ad Hoc Networks 9(3) (May 2011)
14. Zhang, J.: Designing a Cost Effective and Reliable Pipeline Leak Detection System. REL Instrumentation Limited, Manchester, UK

Compressive Data Retrieval with Tunable Accuracy in Vehicular Sensor Networks

Ruobing Jiang[1], Yanmin Zhu[1,2], Hongjian Wang[1],
Min Gao[3], and Lionel M. Ni[1,4]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] Shanghai Key Lab of Scalable Computing and Systems
[3] Guangzhou HKUST Fok Ying Tung Graduate School
[4] Hong Kong University of Science and Technology
{likeice,yzhu,hwang}@sjtu.edu.cn, mingao@ust.hk, ni@cse.ust.hk

Abstract. On-demand data retrieval is a crucial routine operation in a vehicular sensor network. However, on-demand data retrieval in a vehicular environment is particularly challenging because of frequent network disruption, large number of data readings and limited transmission opportunities. Real world vehicular datasets usually contain a lot of *data redundancy*. Motivated by this important observation, we propose an approach called *CDR* with compressive sensing for on-demand data retrieval in the highly dynamic vehicular environment. The distinctive feature of CDR is that it supports *tunable accuracy* of data collection. There are two major challenges for the design of *CDR*. *First*, the sparsity level of the vehicular dataset is typically unknown beforehand. *Second*, it is even worse that the sparsity level of the dataset is changing over time. To combat the challenge posed by time-varying data sparsity, *CDR* can terminate from further collection of measurements, based on an adaptive condition on which only localized measurements and computation are needed. Extensive simulations with real datasets and real vehicular GPS traces show that our approach achieves good performance of data retrieval with user-customized accuracy.

Keywords: vehicular sensor networks, data retrieval, compressive sensing, tunable accuracy.

1 Introduction

Thanks to the rapid advance in embedded senors and inter-vehicle radio communications such as Dedicated Short-Range Communications, vehicular sensor networks (VSNs) [1] [2] have recently attracted growing attention from academy, industry and government. Equipped with on-board sensors such as camera, GPS receiver, and 3D accelerometer, mobile vehicles become powerful mobile sensors. As vehicles may move around in a large area and visit different places, they are able to collect useful sensing data that are geo-distributed vastly. Many applications of vehicular sensor networks have been developed, such as road traffic estimation, road surface monitoring, and proactive urban surveillance [3].

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 426–439, 2013.

A vehicle sensor network consists of a lot of self-organizing mobile vehicles. On each vehicle, an array of useful sensors are embedded onboard. As a vehicle is moving around, the sensors can continuously collect sensing data, e.g., air pollution data. When a vehicle is within the communication range of other vehicles, it can share its sensing data with others. A vehicle that wants to learn about the air pollution distribution in the city can get such sensed air pollution data from other vehicles.

On-demand data retrieval is therefore a crucial routine operation in a vehicular sensor network, with which a querying vehicle retrieves all sensing data of the other vehicles. However, on-demand data retrieval in a vehicular environment is difficult because the following characteristics of vehicular ad hoc networks. *First*, there are a large number of vehicles distributed in a vast area. As a result, retrieving all data from the vehicles is very difficult and costly. *Second*, the whole network is typically disconnected and the network topology is highly dynamic. There is no connected path between each pair of vehicles. *Third*, data transmissions between vehicles can only rely on opportunistic inter-vehicle encounters.

There are some existing studies for improving the performance of data retrieval in vehicular sensor networks. A few studies [4] [5] [6] propose to exploit the fixed infrastructure of roadside units or access points to increase inter-vehicle encounter opportunities and then to improve the data retrieval performance. However, the cost of deploying such a large-scale infrastructures is prohibitively high.

Network coding based approaches [7] [8] have also been proposed for increasing data retrieval performance in vehicular environment. The main idea is to encode original sensed data items into encoded blocks by linearly combining these data items with random coefficients. This enables the quick spread of original sensed data and improves the data accessibility. When the number of collected encoded blocks is close to the number of original sensed data items, the Gaussian elimination technique is adopted to decode all the original data items. However, in order to decode all the sensed data items, a large number of encoded blocks should be collected, which would incur a large retrieval delay and high transmission overhead.

Recently, *compressive sensing* [9] has gained growing importance for its unique capability to recover *redundant* or *sparse* data with only a very limited set of measurements. Mathematically, a dataset is sparse if the number of non-zero values is much smaller than the number of all values in the dataset. A dataset may not be sparse in the original domain. However, the dataset is still sparse as long as there exists a domain in which the transformed dataset is sparse.

Motivated by the observation that real world vehicular datasets are usually sparse, we employ compressive sensing for on-demand data retrieval in the highly dynamic vehicular environment. The main idea is to proactively distribute projected data among vehicles after the original data are generated. A retrieving node can then collect a small set of projected data from those encountered vehicles, with which it estimates the original data with high accuracy by applying the technique of compressive sensing.

Unfortunately, two major challenges must be addressed before the compressive sensing based approach is effective. *First*, the sparsity level of the vehicular dataset is typically unknown beforehand. Second, it is even worse that the sparsity level of the dataset is changing over time. Such time-varying feature of the data sparsity induces great challenges to designing efficient compressive data retrieval approaches.

In this paper, we propose an approach called *CDR* for efficient data retrieval in vehicular sensor networks. *The distinctive feature of CDR is that it supports tunable accuracy of data collection.* To combat the challenges posed by time-varying data sparsity, *CDR* incrementally collects the set of projected data, based on which it recovers the vehicular sensing data. To terminate from further collection of projected data, *CDR* devises an adaptive condition on which only localized measurements and computation are needed. We have performed extensive simulations with real datasets and real vehicular GPS traces. The simulation results show that our approach achieves good performance of data retrieval and can successfully recover vehicular readings with a user-customized accuracy.

The main technical contributions that we have made in the paper are as follows.

- This is the first work, to the best of our knowledge, that deals with the changing sparse level of datasets for compressive sensing based on-demand data retrieval in vehicular sensor networks.
- We have proposed *CDR*, an approach for data retrieval in highly dynamic vehicular environments. With *CDR*, each querying node can retrieve network-wide data retrieval with tunable accuracy. A condition is devised for each querying node to terminate from further collection of projected data, and the condition can be fully evaluated based on localized measurements and computation.
- We have performed extensive simulations based on real vehicular datasets of real vehicular GPS traces, and simulation results show that our approach achieves good performance of data retrieval.

The rest of the paper is organized as follows. The next section reviews related work. Section 3 presents the system model, introduction of basic compressive sensing and problem statement. The basic idea is given in Section 4 and the design details of *CDR* are elaborated in Section 5. The performance evaluation is presented in Section 6. Finally, we conclude the paper in Section 7.

2 Related Work

Previous work on data retrieval in vehicular sensor networks either assume the availability of infrastructures, e.g., road side units [4], or use coding schemes [10] to improve data availability.

CGP [5] takes advantage of road side units (RSUs) to collect datasets from a vehicular environment. The main design consideration is to efficiently use the precious communication chances between RSUs and vehicles. The approach clusters local vehicles on the same road segment and aggregates data in each cluster. Then the aggregated data in each local cluster are relayed by the cluster head to reach the nearest RSU. The local data aggregation reduces the communication between vehicles and road side units. Thus, the precious upload bandwidth can be efficiently used with less collisions.

DB-VDG [6] considers the vehicular data gathering for a base station under specified time constraints. The base station geocasts its data collection query in its local area. The vehicles in the area around the base station that receive the query collect and transmit their sensed data toward the base station. The query generated by the base station includes a specified time interval to limit the delay of data collection process. The data collection process is active only during the specified time interval.

Such infrastructure based data gathering can not provide the data availability for each individual vehicles. Only RSUs or base stations can get the interested dataset from sensing vehicles. They do not support on-demand data retrieval in vehicular networks.

To support data availability for individual vehicles, new approaches are proposed. Roadcast [11] can provide data availability to each individual vehicle. Roadcast explores the popularity of data to ensure that more popular data are more likely to be shared and have more replicas in the vehicular network. However, Roadcast can only provide the most relevant data to the queries of vehicles. Roadcast can not provide vehicles with the whole data sensed by the whole vehicular sensor network.

Various coding schemes, e.g., network coding and erasure coding, are used to provide accessability to the whole data for each individual vehicle. CodeTorrent [7] and VANETCODE [8] use randomized network coding to combine original data packets generated by vehicles into coded blocks. When the number of collected coded blocks is comparable to the number of all the original data packets, the original data packets can be all recovered. However, the relationship between the number of collected coded blocks and the decodability is not deterministic. In another word, the decodability is not guaranteed even when a large number of coded blocks have been collected.

3 Model and Preliminaries

3.1 System Model

The vehicular sensor network comprises a set of vehicles denoted by $V = \{1, 2, \cdots, N\}$. A vehicle $i \in V$ periodically generates a data reading x_i during the period from the area it travels. Any vehicle $v \in V$ may have the demand to retrieve the set of all data readings $\mathbf{x} = [x_1, x_2, \cdots, x_N]^T$. Vehicles in the sensor network communicate with each other when they are within the communication range of each other.

To increase the data retrieval performance in terms of shorter retrieval delay and higher retrieval accuracy, it is a general approach to introduce two processes. More specifically, the two processes are introduced as follows.

- **Data Replication Process.** Firstly, the *data replication process* starts soon after the dataset is generated. In this process, vehicles proactively distribute data replication or redundancy. This process lasts for a fixed time length which is called **replication delay** denoted by α. After this process ends, vehicles cannot generate new data replication or redundancy.

- **Data Retrieval Process.** Next, the *data retrieval process* can be started after the data replication processes terminates. When a retrieving node begins to retrieve the dataset, the data retrieval process is started. The retrieving node tries to recover the original dataset by collecting the replicated data from those vehicles that it has encountered, other than each original source vehicle. The time length of the data retrieval process of vehicle i is called **retrieval delay**, denoted by β_i.

The main objectives of efficient data retrieval are as follows. *First*, for any querying vehicle, the retrieval delay should be as short as possible. *Second*, the retrieved dataset should have a high accuracy, meaning that the retrieved dataset should be as close to the original dataset as possible. *Third*, a lower transmission overhead incurred in the whole network is preferred.

3.2 Basics of Compressive Sensing

Compressive sensing enables a potentially large reduction in the sampling for a sparse signal. A signal is sparse if it can be represented using only a few non-zero coefficients in a suitable basis. Then nonlinear optimization can be used to recover such signal with a few samplings.

For a sensing reading vector $\mathbf{x} = [x_1, x_2, \cdots, x_N]^T \in \mathbb{R}^N$, suppose it is sparse in some transform basis $\Phi = \{\phi_i\}_{i=1}^N$ (e.g., wavelet, Fourier) which is usually orthonormal or orthogonal. Then $\mathbf{x}$ can be represented as the product of Φ^{-1} and a sparse *coefficient vector* $\mathbf{d}$,

$$\mathbf{x} = \Phi^{-1}\mathbf{d} = \sum_{i=1}^{N} \phi_i d_i, \tag{1}$$

in which d_i is the coefficient for the basis vector ϕ_i and $\mathbf{d}$ is sparse in terms that the number of non-zero coefficients in $\mathbf{d}$ is small. Moreover, $\mathbf{x}$ is K-sparse if the number of non-zero coefficients is no more than K, i.e., $\|\mathbf{d}\|_0 \leq K$, in which $\|\mathbf{d}\|_0 := |supp(\mathbf{d})|$. When $K \ll N$, we can recover $\mathbf{x}$ with a small number of measurements.

In compressive sensing, a measurement y is a projection (defined as inner product) on the vector $\mathbf{x}$ with a *projection vector* $\psi = [p_1, p_2, \cdots, p_N]^T$, i.e., $y = \psi^T\mathbf{x} = \sum_{i=1}^{N} p_i x_i$. With M measurements, we have the following equation

$$\mathbf{y} = \Psi\mathbf{x} = [\psi_1^T\mathbf{x}, \psi_2^T\mathbf{x}, \cdots, \psi_M^T\mathbf{x}]^T, \tag{2}$$

where $\Psi = [\psi_1, \psi_2, \cdots, \psi_M]^T$ is an $M \times N$ *measurement matrix.* According to [12], the sensing vector $\mathbf{x}$ can be successfully reconstructed when the matrix Ψ satisfies the restricted isometry property (RIP) of order $2K$. Moreover, the number of measurements M should satisfy the following condition to achieve the RIP given the sparsity level K,

$$M \geq CK \log(\frac{N}{K}), \tag{3}$$

where $C \approx 0.28$.

Given M measurements, the following ℓ_1-norm minimization is solved to construct an estimated $\hat{\mathbf{d}}$ to recover the sparse coefficient vector $\mathbf{d}$,

$$\hat{\mathbf{d}} = \arg\min_{\mathbf{z}} \|\mathbf{z}\|_1, \text{s.t. } \mathbf{y} = A\mathbf{z}, \tag{4}$$

where $A = \Psi\Phi^{-1}$ which is also an $M \times N$ matrix and the ℓ_p norm is defined for $p \in [1, \infty]$ as

$$\|\mathbf{z}\|_p = \begin{cases} \left(\sum_{i=1}^{n} |z_i|^p\right)^{\frac{1}{p}}, p \in [1, \infty); \\ \max\limits_{i=1,2,\dots,n} |z_i|, p = \infty. \end{cases} \tag{5}$$

Finally, the approximation $\hat{\mathbf{x}}$ to $\mathbf{x}$ is constructed with sufficient accuracy when the measurement matrix Ψ holds the RIP,

$$\mathbf{x} = \hat{\mathbf{x}} = \Phi^{-1}\hat{\mathbf{d}}. \tag{6}$$

Note that in practice, we usually use random projection vectors to generate measurements.

4 Basic Idea

We propose to apply compressive sensing to sparse vehicular data retrieval. Our compressive data retrieval (CDR) approach includes two schemes for the two processes introduced in Section 3.1, respectively. In the data replication process, each vehicle shares projections of sensed data with other vehicles once it encounters other vehicles. In such way, the original sensing data generated by each vehicle can be spread over the whole network. In the data retrieval process, a querying vehicle can gather measurements (i.e., data projections) from all the vehicles it encounters. Taking advantage of compressive sensing, the querying vehicle can recover the dataset with only a small number of measurements. The number of measurements is much smaller compared with the size of the dataset.

The key issue of *CDR* is to determine the number of measurements a querying vehicle should take given the user-customized recovery accuracy. In other words, the querying vehicle should determine when to stop the collection of measurements, i.e., the termination condition for the retrieval process. It is challenging because different users may have different demands on the recovery accuracy and retrieval delay. There is a tradeoff between the recovery accuracy and the retrieval delay. According to compressive sensing, more measurements provide higher recovery accuracy. However, in order to improve recovery accuracy by collecting more measurements, a querying vehicle should spend more time. What

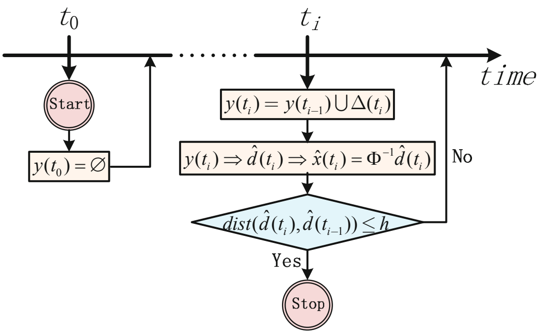

t_0
t_i
time
Start
$y(t_0)=\varnothing$
$y(t_i)=y(t_{i-1})\cup\Delta(t_i)$
$y(t_i)\Rightarrow\hat{d}(t_i)\Rightarrow\hat{x}(t_i)=\Phi^{-1}\hat{d}(t_i)$
No
$dist(\hat{d}(t_i),\hat{d}(t_{i-1}))\leq h$
Yes
Stop

Fig. 1. Illustration of the basic idea. A querying vehicle v raises a data retrieval request at time t_0 and initializes the set of measurements, $\mathbf{y}(t_0)$, as an empty set. At each future time, t_i, when v encounters a set of vehicles, v gathers measurements and decides whether to terminate the data retrieval process.

makes the problem worse is that the sparsity level K of the vehicular dataset is unknown and varies over time. Traditionally, previous approaches usually assume the availability of K and stop the data retrieval process when the number of collected measurements is sufficient (explained in (3)).

Without the knowledge of sparsity level, K, in vehicular environments, we instead evaluate the distance between two sequentially constructed sparse coefficient vectors to decide the termination condition. Because the coefficient vector is the transformed result of the original dataset in another domain, the smaller the distance between two sequentially constructed coefficient vectors, the smaller the distance between two sequentially recovered original datasets. We introduce a threshold of the distance between two sequentially constructed coefficient vectors to adjust the tradeoff between the recovery accuracy and the retrieval delay.

We illustrate the basic idea described above with Fig. 1. Suppose a querying vehicle v queries for the dataset $\mathbf{x}$ at time t_0. v maintains a set of measurements, denoted by $\mathbf{y}$. The set of measurements at t_0 is initialized as an empty set, i.e., $\mathbf{y}(t_0) = \emptyset$.

At each time $t_i > t_0$, the main steps taken by v are as follows. Suppose v encounters a set of vehicles, denoted by $U = \{1, 2, \cdots, u\}, U \subset V$. *First*, it asks each of them to send to it a measurement $y_i, i \in U$. Then v gets a set of measurements $\Delta(t_i)$ from vehicles in U. *Second*, v updates the maintained set of measurements by merging the previous one and $\Delta(t_i)$, i.e., $\mathbf{y}(t_i) = \mathbf{y}(t_{i-1}) \cup \Delta(t_i)$. *Third*, v constructs the sparse vector $\hat{\mathbf{d}}(t_i)$ based on $\mathbf{y}(t_i)$ by solving (4). *Fourth*, i computes the distance between $\hat{\mathbf{d}}(t_i)$ and $\hat{\mathbf{d}}(t_{i-1})$. If the distance is less than a specific threshold h, then v terminates further collections of measurements.

The user-customized threshold h adjusts the tradeoff between the recovery accuracy and the retrieval delay. When a smaller h is specified, i.e., higher recovery accuracy is required by the user v, more measurements are needed, which costs more time.

5 Design of *CDR*

5.1 Overview

In this section, we will introduce the detailed design of our compressive data retrieval (CDR). *CDR* is composed of two main components, namely *data projection scheme* and *data retrieval scheme.*

Data projection scheme is used by each vehicle in the data replication process which specifies how to project sensor readings. We consider the scenario where vehicle i encounters vehicle j at time t and should send a data projection $y_i(t)$ to j. Because j also do the same thing as i, we only consider the actions taken by i. Suppose at time t, the set of projections contained by i is denoted by $\mathbf{y}_i(t)$ and for each $y_s \in \mathbf{y}_i(t)$, the set of sensing data projected in y_s is denoted by $\mathbf{x}_s$.

The main steps are as follows. First, vehicle i selects a subsection $\mathbf{y'}_i(t) = \{y_1, y_2, \cdots, y_r\} \subseteq \mathbf{y}_i(t)$ to make the projection $y_i(t)$. Second, i generates a random coefficient $\psi_i(t)$ for the sensing data x_i. Third, i projects the $y_i(t)$ as follows,

$$y_i(t) = \psi_i(t)x_i + \sum_{s=1}^{r} y_s, y_s \in \mathbf{y'}_i(t). \tag{7}$$

Fourth, i send $y_i(t)$ to j along with the index list of all the sensing data projected in $y_i(t)$ and the corresponding coefficients. Note that the index list of all the sensing data projected in any $y_s \in \mathbf{y'}_i(t)$ and the corresponding coefficients are available because i received them along with the projections y_s.

Initially, when i first encounters a vehicle, it only sends the projection $y_i(t_0) = \psi_i(t_0)x_i$, the index of x_i (i.e., i) and the corresponding coefficient $\psi_i(t_0)$ to the encountered vehicle. We can also find that each $y \in \mathbf{y}_i(t)$ is a projection received by i from other vehicles at time earlier than time t.

There is one key issue for the data projection scheme. To enable the accurate recovery of data vector with high probability for any querying vehicle, the coefficients generated by all the vehicles for their own sensing data should satisfy a condition explained later in Section 5.2.

Data retrieval scheme is designed to decide the termination condition for the retrieval process of a querying vehicle. Suppose the querying vehicle is i and it starts its data retrieval process at time t_0.

Initially, at time t_0, vehicle i initializes its measurement set $\Omega(t_0)$ as empty set and its approximated sensing data vector $\hat{\mathbf{x}}(t_0)$ as zero vector with N elements of zero.

The main steps at time $t > t_0$ are as follows when it encounters one or more vehicles. We only consider one vehicle j that i encounters because each encountered vehicle does the same thing as j. We denote by $\mathbf{y}_z(t)$ the set of projections of vehicle z at time t. *First*, i asks j to send to i a projection of a subset of $\mathbf{y}_j(t)$. *Second*, j makes a projection $y_j(t)$ to send to i using the same projection scheme applied in the data replication process, i.e., the data projection scheme mentioned above. *Third*, i updates the measurement set by adding $y_j(t)$ to $\Omega(t-1)$ to construct the new measurement set $\Omega(t) = \{\omega_s\}_{s=1}^{\pi}$.

Fourth, i approximates the sparse coefficient vector $\hat{\mathbf{d}}(t)$ by solving the following optimization using basis pursuit [13]

$$\begin{aligned} &\hat{d}(t) = \arg\min_{u} \|u\|_1 \\ &\text{s.t. } [\omega_1, \omega_2, \cdots, \omega_\pi]^T = \Psi\Phi^{-1}u, \end{aligned} \tag{8}$$

where $\Psi = [\psi_1^T, \psi_2^T, \cdots, \psi_\pi^T]^T$ is the measurement matrix with ψ_s being the coefficient row vector of sensing data projected in ω_s, and Φ is the Haar wavelet transform basis. Fifth, i evaluates the termination condition by comparing $\hat{\mathbf{d}}(t)$ with $\hat{\mathbf{d}}(t-1)$ and decides whether to collect new measurements or terminate the data retrieval process.

The *key issue* for the data retrieval scheme is to determine the termination condition to balance the tradeoff between the data retrieval delay and the data recovery accuracy.

Next we deal with the key issues in the two schemes.

5.2 Random Gaussian Data Projection

As proved in [14], the $M \times N$ measurement matrix Ψ in (2) holds the RIP [12] with high probability, if the entries ψ_{ij} of Ψ are independent realizations of Gaussian random variables as follows

$$\psi_{ij} \sim \mathcal{N}\left(0, \frac{1}{M}\right), \tag{9}$$

where ψ_{ij} is the entry at ith row and jth column of Ψ.

As a result, in order to make sure the final measurement matrix holds the RIP, the data projection scheme has two rules. First, the coefficients of x_i generated at different time by vehicle i follows the i.i.d. Gaussian random variables as listed in (9). Second, vehicle i selects a subset $\mathbf{y'} = \{y_1, y_2, \cdots, y_r\} \subseteq \mathbf{y}$ to make the projection $y_i(t)$, i.e., $y_i(t) = \psi_i(t)x_i + \sum_{s=1}^{r} y_s$. Suppose the set of sensing data projected in y_s is denoted by $\mathbf{x}_s$. Then the subset $\mathbf{y'}$ should satisfy that any pair of $y_p, y_q \in \mathbf{y'}, \mathbf{x}_p \cap \mathbf{x}_q = \emptyset$. The two rules of our data projection scheme ensures that the coefficients of sensing data in each measurement is the initially generated one, i.e., entries in the measurement matrix of any querying vehicle are i.i.d..

5.3 Termination Condition

The data retrieval scheme solves the issue for each querying vehicle that when to stop the data retrieval process.

Previous work usually assume that the sparsity level K of the sensing data is known. Then a querying vehicle can stop its data retrieval process just when it has collected enough number of measurements given in (3). However, under real environment, we do not have the knowledge of K and the worse thing is that

K varies over time. Instead, our data retrieval scheme observes the sequentially recovered coefficient vector $\hat{\mathbf{d}}$ to decide when to stop with the awareness of the recovery accuracy.

We first explore the relation between the recovery error of $\mathbf{d} = \Phi\mathbf{x}$, denoted by $\varepsilon(\mathbf{d}, \hat{\mathbf{d}})$ and the distance between sequential approximations of $\mathbf{d}$. By $\hat{\mathbf{d}}_M$ we denote the recovered coefficient vector with M measurements.

According to [13], for a sparse vector $\mathbf{d}$ with K non-zero elements, if $\hat{\mathbf{d}}_M + 1 = \hat{\mathbf{d}}_M$ and the entries in the measurement matrix are realizations of a Gaussian random variable, then $\hat{\mathbf{d}}_M = \mathbf{d}$, with probability 1. This proposition holds for any optimization algorithm to the minimization of (4) including basis pursuit which is applied in *CDR*.

As a result, we stop each data retrieval process when the normalized difference ε between two sequentially recovered coefficient vectors is less than a threshold h, i.e., when

$$\varepsilon = \frac{\left\|\hat{d}_M - \hat{d}_{M+T}\right\|_2}{\left\|\hat{d}_M\right\|_2} \leq h. \tag{10}$$

6 Performance Evaluation

6.1 Methodology and Simulation Setup

We conduct extensive trace-driven simulations to evaluate the performance of our compressive data retrieval (*CDR*) with real vehicular sensing datasets. We compare *CDR* with other two data retrieval approaches. The main performance metrics we consider are the data retrieval delay, and data recovery error. The data recovery error is computed as $\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2}$, in which $\hat{\mathbf{x}}$ is the approximation of $\mathbf{x}$.

The datasets used in simulations include a real dataset of vehicular speed readings and a synthetic dataset generated with specified sparsity level K. The real dataset were collected from real traces of 2,600 taxis in urban Shanghai, China in January, 2006. Each taxi is equipped with a Global Positioning System (GPS) receiver which periodically reads vehicular positions and speeds.

Two sets of simulations are conducted. The first set of simulations evaluate the performance of three data retrieval approaches under different system parameters including the number of vehicles, data replication delay α and data retrieval delay β. The second set of simulations focus on the effect of the threshold h on the recovery error and retrieval delay of *CDR* approach.

Vehicles move on the road network of urban Shanghai, China and communicate with each other when they are within the communication range of each other. Each time a vehicle encounters another vehicle, only one data projection can be transmitted.

For each simulation setup, 15 runs are conducted and the average results are plotted. The default system parameters are shown in Table. 1.

Table 1. Default System Parameters in Simulations

Parameter	Default Value
Communication Range	150 m
Number of Vehicles	1000
Replication Delay α	4 hours
Retrieval Delay β	2 hours
Threshold h	0.15
Sparsity Level K (synthetic datasets)	10

6.2 Compared Approaches

The two compared data retrieval approaches are two combinations of a simple replication scheme SR and two representative estimation schemes, $k-$NN estimate (KNN) and Gaussian process regression (GP).

- Simple replication (SR). In the data replication process, each vehicle transmits the original sensing data to other vehicles. If a vehicle i has more than one sensing data and encounters another vehicle j, it randomly selects and transmits to j a sensing data that j does not have.
- $k-$NN estimate (KNN). To better recover sensing data, a querying vehicle sets the value of a unavailable sensing data x_i as the average of all the sensing data from k nearest neighboring vehicles of i.
- Gaussian process regression (GP). It interpolate values of a random field at unobserved locations from observations of its value at nearby locations.

The two compared approaches are SR+KNN and SR+GP.

6.3 Performance Comparison

We compare the performance of three approaches under different system parameters for vehicular speed dataset and synthetic dataset.

First, when the number of vehicles varies from 200 to 1,000, the recovery error of all the approach are shown in Fig. 2 and Fig. 5 for speed dataset and synthetic dataset, respectively. *CDR* has the lowest error (lower than 0.4 for speed dataset and lower than 0.2 for synthetic dataset) while the recovery error of SR+KNN and SR+GP stay above 0.9 for both datasets. We can find that when the number of vehicles is extremely small (e.g., 200), the recovery error of *CDR* is much higher than those situations where more vehicles exist. This is because that when less vehicles exist, the encounter chances among vehicles in the network are much lower. Thus, vehicular data are difficult to spread around the whole network which results in the relatively higher recovery error.

Second, when the replication delay α increases from 1 hour to 5 hours, Fig. 3 and Fig. 6 plot the recovery error of all the approaches for different datasets, respectively. For both datasets, *CDR* performs best. In Fig. 3, the recovery error of *CDR* decreases from 0.6 to 0.2. For synthetic dataset, the recovery error of *CDR* decreases below 0.1 when the replication delay is longer than 3 hours.

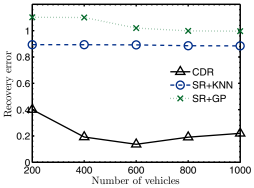


Fig. 2. Recovery error vs. number of vehicles (speed dataset)

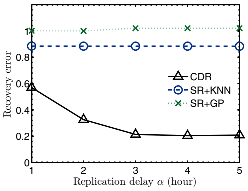


Fig. 3. Recovery error vs. replication delay (speed dataset)

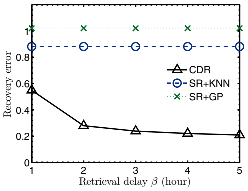


Fig. 4. Recovery error vs. retrieval delay (speed dataset)

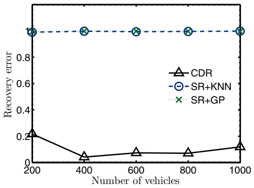


Fig. 5. Recovery error vs. number of vehicles (synthetic dataset)

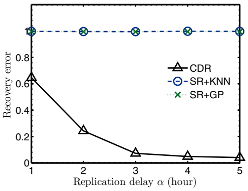


Fig. 6. Recovery error vs. replication delay (synthetic dataset)

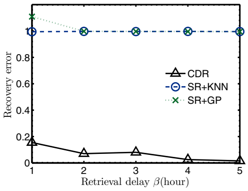


Fig. 7. Recovery error vs. retrieval delay (synthetic dataset)

Third, we present the recovery error of three approaches when the retrieval delay increases from 1 hour to 5 hours in Fig. 4 and Fig. 7. All the compared approaches have lower recovery error as the retrieval delay increases. Moreover, we can find that the recovery error of *CDR* for the synthetic dataset is lower than that for the speed dataset under same retrieval delay. This is because the real speed dataset is not strictly sparse while the synthetic dataset is strictly sparse. There is noise for datasets which is not strictly sparse which results in higher recovery error.

For SR+KNN and SR+GP, the reason why their recovery error keep higher than 0.9 is that the scarce transmission chances can not satisfy the traffic load for all the sensing data to spread over the whole network. By taking advantage of sparsity property, *CDR* can achieve better performance.

6.4 Effect of h

We then explore the effect of threshold h on the performance of *CDR* for a synthetic dataset. Specifically, we evaluate the recovery error, number of collected measurements, and retrieval delay of *CDR*. The sparsity level of the synthetic dataset is set as 10. The threshold h varies from 0.05 to 0.25. Each plotted result is an average value of 15 runs of simulations.

Fig. 8 plots the recovery error of *CDR*. As expected, the recovery error decreases when smaller h is specified. We can also find that the recovery error is

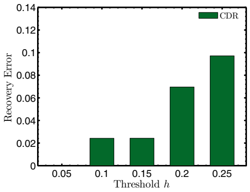


Fig. 8. Recovery error vs. threshold h (synthetic dataset)

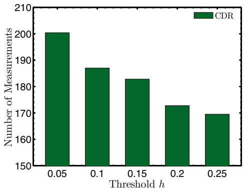


Fig. 9. Number of measurements vs. threshold h (synthetic dataset)

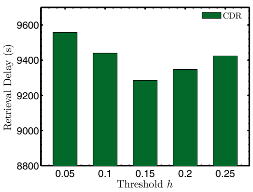


Fig. 10. Retrieval delay vs. threshold h (synthetic dataset)

smaller than h. As a result, user-customized recovery accuracy can be achieved by adjusting the parameter h.

Fig. 9 presents the number of collected measurements when different threshold, h, is specified. As expected, the number of measurements decreases as the bigger h is set. It is reasonable because bigger h means lower recovery accuracy is required which needs less measurements. We can also find that the relation between the number of measurements and the value of h is approximately linear.

Fig. 10 shows the data retrieval delay under different threshold. From the figure, we can not find clear relation between the retrieval delay and h. This is mainly because that the retrieval delay is not only determined by specified recovery error, but also related to the encounter rate between the querying vehicle and other vehicles. When the recovery error is specified or a specified h is given, the number of measurements is also determined. However, the delay to collect a given number of measurements varies for different vehicles. If the vehicle is located in a dense area with high encounter rate with other vehicles, then the delay will be much small. On the other hand, when the vehicle is located in a sparse area with much less vehicles, then the delay to collect a fixed number of measurements will be much longer.

7 Conclusion

In this paper we have focused on the crucial problem of on-demand data retrieval in a highly dynamic but challenged vehicular environment. Inspired by the observation that real vehicular datasets are usually sparse, we have proposed *CDR* for on-demand data retrieval with tunable accuracy. With compressive sensing, *CDR* realizes tunable accuracy by devising a condition on which a retrieving node can test based on local measurements and stops further collection of measurements. This effectively combats the challenge posed by time-varying sparsity of the vehicular sensing datasets. Based on the real vehicular sensing datasets and real vehicular GPS traces collected from around 2,600 taxis in Shanghai, China, our simulation results demonstrate that *CDR* achieves good performance of data recovery accuracy.

Acknowledgements. This research is supported by NSFC (No. 61170238, 60903190, 61027009, 60933011, 61202375, 61170237), Shanghai Pu Jiang Talents Program (10PJ1405800), Shanghai Chen Guang Program (10CG11), MIIT of China (2009ZX03006-001-01), Doctoral Fund of Ministry of Education of China (20100073120021), National 863 Program (2009AA012201 and 2011AA010500), HP IRP (CW267311), SJTU SMC Project (201120), STCSM (08dz1501600, 12ZR1414900), Singapore NRF (CREATE E2S2), and Program for Changjiang Scholars and Innovative Research Team in Universities of China (IRT1158, PCSIRT).

References

1. Lee, U., Gerla, M.: A survey of urban vehicular sensing platforms. Computer Networks 54(4), 527–544 (2010)
2. Zhang, C., Lu, R., Lin, X., Ho, P.H., Shen, X.: An efficient identity-based batch verification scheme for vehicular sensor networks. In: Proc. IEEE INFOCOM, pp. 246–250. IEEE (2008)
3. Lee, U., Zhou, B., Gerla, M., Magistretti, E., Bellavista, P., Corradi, A.: Mobeyes: smart mobs for urban monitoring with a vehicular sensor network. IEEE Wireless Communications 13(5), 52–57 (2006)
4. Yang, L., Xu, J., Wu, G., Guo, J.: Road probing: Rsu assisted data collection in vehicular networks. In: 5th International Conference on Wireless Communications, Networking and Mobile Computing, WiCom 2009, pp. 1–4. IEEE (2009)
5. Salhi, I., Cherif, M.O., Senouci, S.M.: A new architecture for data collection in vehicular networks. In: IEEE International Conference on Communications, ICC 2009, pp. 1–6. IEEE (2009)
6. Palazzi, C.E., Pezzoni, F., Ruiz, P.M.: Delay-bounded data gathering in urban vehicular sensor networks. Pervasive and Mobile Computing 8(2), 180–193 (2012)
7. Lee, U., Park, J.S., Yeh, J., Pau, G., Gerla, M.: Code torrent: content distribution using network coding in vanet. In: Proceedings of the 1st International Workshop on Decentralized Resource Sharing in Mobile Computing and Networking, pp. 1–5. ACM (2006)
8. Ahmed, S., Kanhere, S.S.: Vanetcode: network coding to enhance cooperative downloading in vehicular ad-hoc networks. In: Proceedings of the 2006 International Conference on Wireless Communications and Mobile Computing, pp. 527–532. ACM (2006)
9. Baraniuk, R.G.: Compressive sensing (lecture notes). IEEE Signal Processing Magazine 24(4), 118–121 (2007)
10. Fujimura, A., Oh, S.Y., Gerla, M.: Network coding vs. erasure coding: Reliable multicast in ad hoc networks. In: IEEE Military Communications Conference, MILCOM 2008, pp. 1–7. IEEE (2008)
11. Zhang, Y., Zhao, J., Cao, G.: Roadcast: a popularity aware content sharing scheme in vanets. ACM SIGMOBILE Mobile Computing and Communications Review 13(4), 1–14 (2010)
12. Davenport, M.A., Duarte, M.F., Eldar, Y.C., Kutyniok, G.: Introduction to compressed sensing. Preprint 93 (2011)
13. Malioutov, D., Sanghavi, S., Willsky, A.: Compressed sensing with sequential observations. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008, pp. 3357–3360. IEEE (2008)
14. Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices. Constructive Approximation 28(3), 253–263 (2008)

Enforcing Spectrum Access Rules in Cognitive Radio Networks through Cooperative Jamming

Yantian Hou and Ming Li

Department of Computer Science
Utah State University, Logan, Utah 84341

Abstract. In Cognitive Radio Networks (CRNs) with dynamic spectrum access, it is of paramount importance to ensure the spectrum access rules are honestly followed by each secondary user. Existing approaches either require significant modifications to the system hardware, or can only deter the spectrum abuse from happening by punishing abusers after-the-fact, which is ineffective in reality as there lacks universal identification for each device. In this paper, we propose a novel spectrum access rule enforcing scheme by introducing a "spectrum guardian", who punishes abusers immediately on-the-scene through optimally jamming their signals using multiple antennas while without affecting the communication between primary users, thus removing abusers' incentive to exploit the spectrum for their own benefit. Our scheme requires no modifications to Commercial-Off-The-Shelf (COTS) CR devices, nor the need of device identification.

1 Introduction

Cognitive Radio Networks (CRN) has been shown as a promising technology to solve the spectrum under-utilization problem. Dynamic Spectrum Access (DSA) is a key component in CRN system design, which can improve the spectrum efficiency by allowing secondary users (SUs) to use the spectrum as long as the primary/licensed users (PUs) are not using it.

However, in reality the licensed users may not be willing to share the spectrum with SUs, due to both technical and policy barriers. On the one hand, how to sense the spectrum accurately to give PU the transmission priority has been extensively studied. On the other hand, to overcome the policy barriers, incentive mechanisms are needed to encourage opportunistic spectrum sharing. For example, spectrum auction schemes have been proposed to allocate the unused spectrum to SUs [7,8], which maximizes the PU's economic gains. Unfortunately, those incentive mechanisms do not guarantee that the spectrum access rules are correctly followed by every SU. In fact, there exists selfish users who may disobey the spectrum access rules to gain benefits. This is called as the *spectrum abuse* problem, which can become a critical threat to practical adoption of CRN.

Therefore, enforcing mechanisms are needed to stop the spectrum abuse in CRN and guarantee the legitimate spectrum usage. There are several previous

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 440–453, 2013.

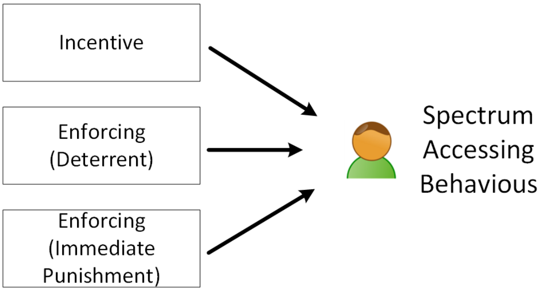

Incentive
Enforcing (Deterrent)
Enforcing (Immediate Punishment)
Spectrum Accessing Behavious

Fig. 1. Three main mechanisms used to deal with spectrum accessing problem. The existing methods can be categorized into incentive or deterrent mechanisms. Our objective is to find an immediate punishment scheme.

works in this direction, but all of them are *detection and deterrence* mechanisms. Sahai et. al's work [15] requires hardware modifications to embed an unique identifier for each CR device, so that abusers can be identified and punished afterwards. However, it is not compatible with COTS devices. Some other researchers [19] propose to embed some credential information in physical layer and rely on patrolling police devices to detect spectrum abuse, or charge a fine to the abuser's account as a punishment. Yet, these approaches still rely on unique device identification, and can only apply the punishment after-the-fact.

Thus, we argue that an actual spectrum access *enforcement* mechanism is still lacking. The main challenge is that, in reality: 1) It is difficult to identify each CR device without hardware modifications. Devices can be self-developed or running different protocols, or can change their MAC addresses. Hardware-level device fingerprinting still requires high cost. 2) The devices could be highly mobile, therefore are very hard to localize. The abusers are likely gone before they are caught. 3) Even worse, the abusers could be *myopic*, which means they only care about short-term gains and not their long-term gains and reputation. Therefore, we ask this question: is there such a mechanism that can stop and punish the abusers immediately after the spectrum abuse happens no matter what type of devices they use nor how uncertain their locations are?

In this paper, we employ the concept of cooperative jamming to achieve this goal. Essentially, we propose to add a "spectrum guardian", who jams an abuser's signal whenever it happens. However, this must be done without affecting legitimate PU's transmission, while considering the abusers' uncertain locations and channel conditions. Thus, we propose a beamforming-based optimized jamming framework (using multiple antennas), where the expected jamming power sensed by the abusers is maximized. Different from existing deterrent-based mechanisms, our method is an immediate punishment based enforcement, as shown in Fig. 1. Our scheme doesn't need any modification to existing CR devices and is compatible with legacy systems.

Our main contributions are: 1) We propose a real spectrum access enforcement approach based on cooperative jamming, which needs no modification to existing CR devices. 2) We formulate an optimized cooperative jamming problem, which considers the uncertainty of channel condition and abuser locations. 3) We evaluate its security level and effectiveness using extensive simulations.

This paper is organized as follows. In section 2, the related works are presented. In section 3, we will introduce our motivation. After that we will introduce the system and attacker model. Then we will introduce the problem formulation in section 4. The simulation results are shown in section 5 and after that is the conclusion.

2 Related Work

2.1 Spectrum Misuse Deterrent

Some works have already considered the spectrum access enforcing problem. In [19], the authors proposed an abuse/misuse deterrent scheme. In their method, each legitimate user should send packets using a spectrum access permit which is embedded in the physical layer. The police devices would patrol in the CRN area to detect spectrum abuse by checking the permit on the air, thus have a deterrent to spectrum abusers. Once the abuser is detected, some localization methods are used to locate the attacker and kick it out from the CRN. Another spectrum access enforcement scheme is proposed in [1] [15] . Their scheme needs identification fingerprint embedded in physical layer of each device. Besides, their scheme requires that all devices must obey a 'go-to-jail' command to enforce their punishment once they are caught in spectrum abusing, and thus it forms a deterrent to spectrum abuse. Even though the authors claim that their scheme is light-handed, it still requires modification of user devices. These two deterrent methods both rely on some identification or permit information on each device, thus cannot deal with attackers which care less about their reputation in the CRN and highly-mobile. Besides, they both need modification on existing devices. The authors in [17] proposed an Ally Friendly Jamming scheme which can disable unauthorized communication while allowing the authorized devices to communicate. However, their scheme relies on secret key equipped by authorized users, which also need modification on the devices.

2.2 Cooperative Jamming

The work closest to our scheme is proposed in [3], [4] which also uses the technic of beamforming from multiple antennas. However, their scheme assumes that all channels' conditions are deterministic, which seems impractical in reality, e.g. the channels' conditions from jammer to attackers are hard to measure. Instead, we consider the uncertainty of the channels to multiple abusers' locations. The authors in [12] considered a similar secret communication problem and then proposed a scheme based on artificial noise. They considered the uncertainty of

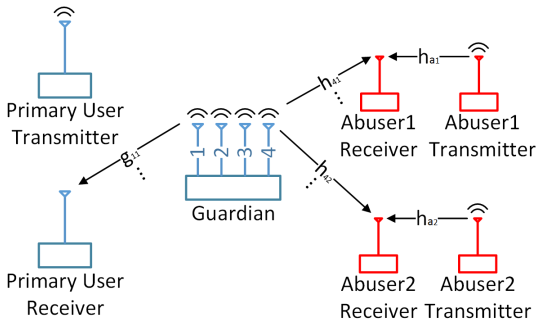


Fig. 2. Illustration of spectrum guardian system

channel to eavesdroper and use a brute force method to solve their problem. However, this paper considered only one eavesdroper in their problem formulation. Besides, their method needs modification on the signals sent by transmitter.

2.3 Interference Cancellation

In [5] the authors proposed TIMO based on channel measurement which can communicate in the presence of interference. However, their scheme only considered cross-technology interference such as cordless phone but cannot deal with multi-antenna jammer. In another word, their scheme could deal with attacker which has only one interference source. The authors in [16] proposed a counter-jamming scheme using mechanical and software interference cancellation. However their scheme only works for static attackers and cannot handle multi-antenna jammers either.

3 Problem Statement

3.1 System Model

In our system, we assume one primary user with static location. The abusers could have several possible locations. We assume the abusers could be at any location with same probability. A multiple-antenna device is deployed in our system as the guardian. We assume the channel conditions from our guardian to the primary user's receiver could be measured. However, the channel conditions from guardian to abusers are uncertain. We use a path loss and rayleigh fading model for the channels from guardian to abusers. The system model is shown in Fig. 2.

3.2 Attack Model

In this paper, we assume the abusers are selfish but rationale, i.e. they transmit at any time they want, without regarding any spectrum accessing rules, but they care about their own cost, such as energy consumption. The main difference of our abuser model from others [15], [19] is that we consider our abuser to be myopic and highly mobile, i.e. they don't care about their reputation in the CRN and are very hard to be localized. Therefore the deterrent mechanisms are not effective to deal with these abusers. However, we don't consider malicious attacks such as Denial of Service (DOS) attack.

3.3 Design Objective

We assume the abusers could transmit at any time without obeying the spectrum accessing rule. Therefore they could cause interference to the primary user or other legitimate secondary users. We assume the spectrum abuse behaviours could be detected using some existing spectrum sensing methods, such as energy detection and feature detection.

Our goal is to use the multi-antenna device to jam the abusers' transmissions, while causing no interference to the legitimate user. As the abusers could be at several possible locations, we need to generate jam signals at all these locations. We also want the interference generated on all the possible locations to be maximized.

4 Spectrum Access Enforcing

In this section, we will introduce our spectrum access enforcing scheme based on cooperative jamming. We will study the cooperative jamming problem considering the random channel condition and multiple abuser positions. We will formulate this problem as a non-convex problem and solve it using empirical methods, and also compare the results with a lower bound derived from the dual problem.

4.1 Design Overview

We will first introduce our spectrum access enforcing method to deal with spectrum abuse. In our scheme, we don't rely on any identification based methods as these methods cannot deal with myopic users' abuse. Our scheme is an enforcing method, which can generate interference signals to block the abuser's transmission immediately when it happens. A good analogy could illustrate our idea: The traditional identification based abuse deterrent method is like fingerprint detection in the crime scene in public safety issues. It is good deterrent to most rationale, pre-identified people. However it is slow and cannot deal with the people without being identified beforehand. Our method is like police patrolling on the street and stop the crime immediately when it is happening, thus it is fast enforcing method and doesn't rely on any knowledge about the criminal's identity. The only inaccurate aspect in this analogy is that in CRN the spectrum abuse detection could be much easier and more accurate than patrolling on street to search for crimes.

Our spectrum access enforcing method has two modes: The first is reactive jamming mode, in which the guardian starts to jam the abusers when it detects spectrum abuse during primary user's transmission. The second is constant jamming mode, in which the guardian jam whenever the primary user is using the spectrum. The only difference of the two modes is that the reactive mode needs a spectrum abuse detection step. However, the reactive mode consumes less energy than the constant jamming mode. We will introduce the reactive mode in the following.

The reactive jamming mode has three main steps. The first step is abuse detection. The goal is to detect any spectrum abuse when primary user is transmitting. There are several existing spectrum detecting methods can be applied directly, such as energy detection, feature detection, and location based detection [10] [11].

After the abuse detected, we measure the channel from guardian to the primary users' receivers. The guardian first sends a probe signal s to each primary user's receiver. After receiving the probe signal, the receiver will send back a response packet, which contains the received signal s'. Then the channel from guardian to primary user is calculated by the guardian as:

$$g = \frac{s'}{s} \tag{1}$$

After the channel from guardian to primary user's receiver is calculated, the guardian will jam the abusers without generating interference to the primary user. In the following subsection we will consider the uncertainties of abusers' locations and channel conditions in our jamming problem.

4.2 Preliminary on Cooperative Jamming

In cooperative jamming, once the primary user is transmitting, the multi-antenna guardian will send a jamming signal $\mathbf{w}s$. The signals received by primary receiver and abuser can be denoted as $\mathbf{w}^\dagger\mathbf{g}s$ and $\mathbf{w}^\dagger\mathbf{h}s$. $\mathbf{g}$ and $\mathbf{h}$ denote the channels from guardian to the primary receiver and abuser respectively. Our goal is to choose each antenna's weight $\mathbf{w}$ to minimize the abuser's transmission rate while generating no interference to the primary receiver.

First we consider only one abuser location. According to Shannon Capacity Theory, the abuser's transmission rate is bounded by:

$$R_a = \log(1 + \frac{P_s \cdot |h_a|^2}{|\mathbf{w}^\dagger\mathbf{h}|^2 + \sigma^2}) \tag{2}$$

where P_s is the abuser's transmission power, h_a denotes the channel from abuser's transmitter to receiver. σ^2 is the background noise. We can find that minimizing equation 2 is equivalent to maximizing the interference to abuser $|\mathbf{w}^\dagger\mathbf{h}|^2$. Therefore we can simplify our objective function 2 into the equivalent form. Considering our zero interference to primary user and the antenna power constraint, our problem could be formulated as:

$$\begin{aligned} \text{maximize} \quad & |\mathbf{w}^{\dagger}\mathbf{h}|^2 && (3) \\ \text{subject to} \quad & \mathbf{w}^{\dagger}\mathbf{g} = 0 && (4) \\ & \mathbf{w}^{\dagger}\mathbf{w} \leq P_0 && (5) \end{aligned}$$

The constraint 4 denotes that the guardian generates zero interference to primary user's receiver. In equation. 5, the P_0 is the guardian's transmission power constraint, which means the total jamming power should not exceed a power limit P_0.

We should note that in order to satisfy the constraint in equation 4, i.e. to generate zero interference to the primary users, we need the number of antennas of guardian to be not less than the number of primary users. The solution of this problem is given in [3].

4.3 Beamforming Based Optimized Jamming Framework

Problem Formulation. In reality, the channel conditions from guardian to abusers are very likely to be unknown. Therefore we estimate the channel based on propagation models. We use the path-loss and fading model for these channels. According to [13], we model the channel from guardian to abuser as $\alpha e^{j\phi}$, in which α is Rayleigh-distributed envelope, and ϕ is phase randomly distributed within $[-\pi, \pi]$. Besides, there could be multiple abusers, and their locations are uncertain. We consider the abusers could be at any 1 of N locations with the same probability. Thus our objective is to minimize the expectation of the average of abusers' capacities at these possible locations, which could be expressed as:

$$R'_a = E[\frac{1}{N} \cdot \sum_{i=1}^{N} \log(1 + \frac{P_{si} \cdot |h_{ai}|^2}{|\mathbf{w}^{\dagger}\mathbf{h_i}|^2 + \sigma^2})] \tag{6}$$

In equation 6, E denotes the expectation. $\mathbf{h_i} = [h_{1i}h_{2i}...h_{Mi}]^{\dagger}$ is the channel from guardian's all M antennas to ith abuser's receiver. This problem is hard to solve, therefore we consider a suboptimal problem. Instead of minimizing abuser's capacity, we try to maximize the interference to abuser caused by jamming, which can be formulated as:

$$R''_a = E[\frac{1}{N} \cdot \sum_{i=1}^{N} |\mathbf{w}^{\dagger}\mathbf{h_i}|^2] \tag{7}$$

We can transform the equation 7 to get rid of the term E:

$$
\begin{aligned}
E[\frac{1}{N}\cdot\sum_{i=1}^{N}|\mathbf{w}^{\dagger}\mathbf{h_i}|^2] &= E[\frac{1}{N}\cdot\sum_{i=1}^{N}|\sum_{k=1}^{N} w_k\cdot h_{ki}|^2] && (8)\\
&= E[\frac{1}{N}\cdot\sum_{i=1}^{N}|\sum_{k=1}^{M}|w_k|\cdot|h_{ki}|\cdot e^{j(\theta_k+\phi_{ki})}|^2] && (9)\\
&= E[\frac{1}{N}\cdot\sum_{i=1}^{N}(\sum_{k=1}^{M}|w_k|^2\cdot|h_{ki}|^2)] && (10)\\
&= \frac{1}{N}\cdot\sum_{i=1}^{N}(\sum_{k=1}^{M}|w_k|^2\cdot h_{ki}^{'2}) && (11)
\end{aligned}
$$

In 11, the $h_{ki}^{'2}$ denotes the expectation of $|h_{ki}|^2$, which is the average attenuation of channel h_{ki}. In equation 9, θ_k and ϕ_{ki} are the phases of w_k and h_{ki} respectively. We know multi-path fading effect is a relatively small effect compared with path-loss [6], [9]. Therefore we can use the path-loss model to derive $h_{ki}^{'2}$. We transform equation 11 to a matrix multiplication form and then formulate our problem as:

$$
\begin{aligned}
\text{minimize} \quad & -\frac{1}{N}\cdot\sum_{i=1}^{N}\|\mathbf{w}^{\dagger}\cdot\mathbf{H_i}\|^2 && (12)\\
\text{subject to} \quad & \mathbf{w}^{\dagger}\mathbf{g}=0 && (13)\\
& \mathbf{w}^{\dagger}\mathbf{w}\le P_0 && (14)
\end{aligned}
$$

In equation 12, the $\mathbf{H_i}$ is $diag\{\mathbf{h_i}\}$. We can see our problem is not a convex optimization problem. In the next section, we will solve it using Genetic Algorithm (GA). Before that, we will show the lower bound derived from the dual problem.

Lower Bound Analysis. We solve the dual problem of our original problem to get a lower bound of it. We first rewrite equation 12 to a equivalent form $-\frac{1}{N}\cdot\mathbf{w}^{\dagger}\mathbf{H}^{*}\mathbf{w}$, in which $\mathbf{H}^{*}$ is the summation $\mathbf{H}^{*}=\sum_{i=1}^{N}\mathbf{H_i}\mathbf{H_i^{\dagger}}$. Then based on this new objective function and constraints 13 and 14, we have the Lagrangian:

$$L(w,\lambda,\nu)=-\frac{1}{N}\mathbf{w}^{\dagger}\mathbf{H}^{*}\mathbf{w}+\lambda(\mathbf{w}^{\dagger}\mathbf{w}-P_0)+\nu(\mathbf{w}^{\dagger}g) \tag{15}$$

Then the dual function is given by:

$$g(\lambda,\nu)=-\frac{1}{4}\nu^2\cdot\mathbf{g}^{\dagger}(\lambda\mathbf{I}-\frac{1}{N}\mathbf{H}^{*})^{-1}\mathbf{g}-\lambda P_0 \qquad \lambda\mathbf{I}-\frac{1}{N}\mathbf{H}^{*}\succeq 0 \tag{16}$$

We can see that in the dual function in equation 16, as the matrix $\lambda\mathbf{I}-\frac{1}{N}\mathbf{H}^{*}\succeq 0$, the left component $-\frac{1}{4}\nu^2\cdot\mathbf{g}^{\dagger}(\lambda\mathbf{I}-\frac{1}{N}\mathbf{H}^{*})^{-1}\mathbf{g}$ must be ≤ 0. Therefore in order to maximize the dual function we should choose $\nu = 0$. Besides, for λ we should choose $\lambda \geq \frac{1}{N}\lambda_{max}$, in which λ_{max} is the maximum eigenvalue of $\mathbf{H}^{*}$, which is $max\{\sum_{i=1}^{N} h_{ki}^2, k = 1, 2, ...M\}$. Thus the dual optimal is $-\lambda_{max} \cdot P_0$, which means the overall interference generated at all the possible abuser locations is up-bounded by $\lambda_{max} \cdot P_0$.

Empirical Solution. As our original problem is non-convex, we use an empirical method to solve it. Here we choose Genetic Algorithm (GA). The population size is set as 20 and the maximum number of iterations is set as 100.

5 Evaluation

In this section, we will find the optimal value and solution of our optimization problem using GA and then compare the derived interference power with the bound derived by the dual problem. We will also verify the correctness of the solution by calculating the interference to the primary users using the derived antennas' weights. The simulation platform is Matlab and the GA algorithm is provided by the optimization tool package in Matlab.

5.1 Experiment Methodology

We choose 5 antennas for the guardian and each antenna is 0.5 meters away from each other. The layout is shown in Fig. 3. We assume an suburban environment and choose the path-loss exponential as 3 [14]. In our simulation we assume the abuser uses 802.11af devices (Note that in reality abuser could choose other devices). The abuser's transmitter-receiver distance is assumed to be 20 meters and the transmitting power is 100 mw [18]. The guardian-primary user's channel parameters are randomly chosen based on the propagation model.

Our evaluation has two parts. In the first part we evaluate the security provided by our scheme. We use simulation method to find the relation between generated interference and abuser distance. Then we analyze the effective jamming range based on the derived relation. In the second part we study the relation between the generated interference and the number of antennas, which could guide our guardian design in real networks.

5.2 Security Analysis

We consider a primary user's receiver which could tolerate interference level at about -93 dBm [2]. Assuming the abuser uses a 100 mw transmitter, we calculate that the maximum interference distance for primary user is 685 meters based on the propagation model. We will study the jamming effect of our multi-antenna guardian within this range using simulation.

We analyze the relation of interference and abuser locations. The layout of this experiment is shown in the left subfigure of Fig. 3. In this part we select one

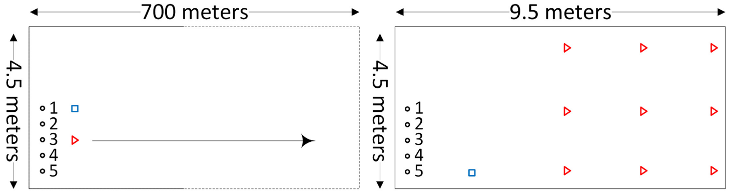


Fig. 3. The layout of our simulations. The circle ○ denotes the guardian's antenna. The square □ denotes the primary user. The triangle △ denotes the abuser. The left figure is for the evaluation part 1 and the right one is for part 2. In left figure, the arrow denotes the moving direction of abuser.

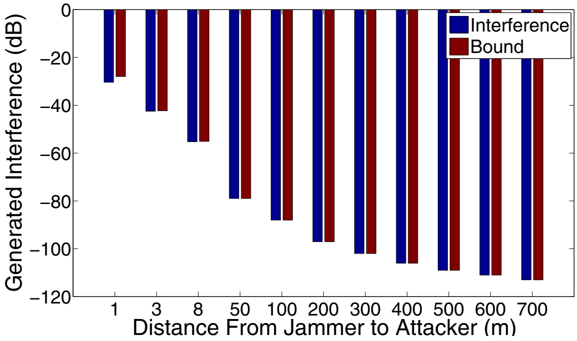


Fig. 4. The Comparison of generated interference power on the abuser at different distances

attacker location for simplicity. We assume the guardian is near the primary user. We choose 11 different distances from guardian to abuser. The distance here is defined from the middle of the guardian (which is the antenna 3 in the middle) to the abuser. We derive the solution and the corresponding interference using GA under each distance from 1 meter to 700 meters. The generated interference to the abuser is shown in Fig. 4.

In Fig. 4, for each location, the left blue column denotes the interference derived by GA. The right red column denotes the bound calculated from the dual problem. In this figure, we can find that as the abuser moves further away from guardian, the interference decreases. When the distance from guardian to abuser is small, e.g. 1 meter as shown at abuser location 1, the interference is high (which is nearly -30 dB). When the distance is 200 meters (at location 6), the interference is still about -95 dB, which is still a high interference. Assuming the abuser device's sensing threshold as -94 dB [18], the result shows our guardian could still jam the abuser at this distance using only 1 w. When the distance is 700 meters, the interference is -113 dB. We could jam the abuser at this distance if we increase the jamming power to 100 w.

We can also compare the interference power derived by GA with the bound. In Fig. 4, it can be observed that the interference derived by GA is always smaller or equal to the bound, which verified the correctness of this bound. Besides, another phenomenon can be observed: as the distance between guardian and abuser becomes larger, the interference derived by GA and the bound become nearly the same.

The reason is that when the distance is small, the channel conditions from all antennas to the abuser could be largely different because the distance from each antenna to the abuser could be very different from each other. Thus the weights on the antennas could largely affect the interference to abuser. However, if the abuser is far away from the guardian, then the channel's conditions from all antennas to the abuser are nearly the same. Therefore the interference to abuser can be determined approximately by the average distance from guardian to abuser.

For the bound, from the expression of dual optimal value in last section, we know it is determined by the largest channel gain, which can also be approximately determined by the average distance from guardian to abuser when the distance between them is large. Thus the bound and interference are nearly the same for large guardian-abuser distance, which means we can use the bound derived by dual problem to estimate the generated interference to abuser under this condition, thus the problem could be simplified.

We can verify the interference power to the primary user. According to our constraint, the antennas should generate zero interference to the primary user through beamforming. We use the antennas' weights derived from GA to calculate the interference to primary user. The results are shown in Fig. 5. From this figure we can see that the interferences from guardian to the primary user are all smaller than -125 dB, which is less than the primary user's interference threshold. Thus the guardian doesn't cause interference to the primary user in these cases.

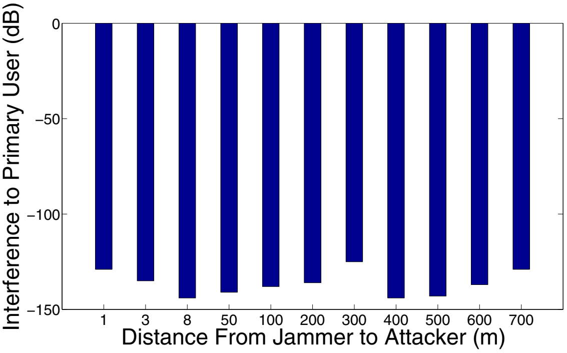


Fig. 5. The interference power to the primary user's receiver when the abuser is at different locations

5.3 Analysis on Number of Antennas

Next we evaluate the relation between the number of antennas and the interference generated to abusers. In this simulation we consider multiple abusers shown in Fig 3. Theoretically the more antennas we have, the more degrees of freedom we could leverage, then the better optimal values we could get. In our simulation, we can see that due to the layout of antennas in our setting, the interference bounds remain the same as the number of antennas increases,

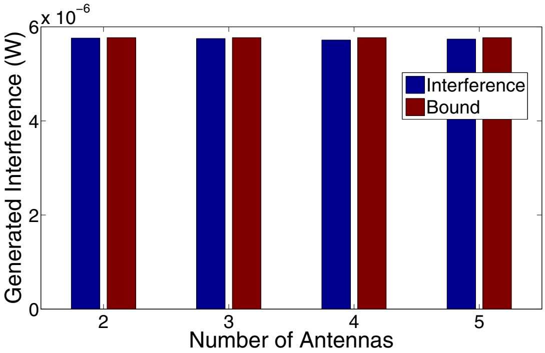


Fig. 6. The Comparison of generated interference power at the abusers with different number of antennas

Table 1. The optimal antennas' weights derived by Genetic Algorithm

Antenna	$No.1$	$No.2$	$No.3$	$No.4$	$No.5$
$2antennas$	$0.7691e^{j4.7644}$	$0.6391e^{j2.2512}$	-	-	-
$3antennas$	$0.7935e^{j6.088}$	$0.6005e^{j3.4134}$	$0.0991e^{j4.2264}$	-	-
$4antennas$	$0.6050e^{j3.7133}$	$0.6709e^{j0.4688}$	$0.3983e^{j1.9866}$	$0.1587e^{j0.4381}$	-
$5antennas$	$0.9445e^{j6.2832}$	$0.1309e^{j3.4038}$	$0.1948e^{j3.4187}$	$0.1794e^{j0.0017}$	$0.1436e^{j1.5344}$

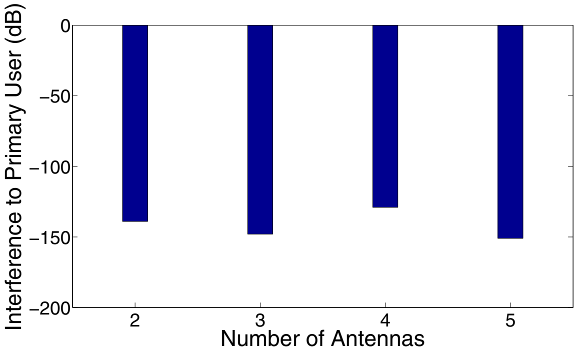


Fig. 7. The interference power to the primary user's receiver using different number of antennas

which can be observed from Fig. 6. We find the interferences derived by the GA are slightly different (the interferences using 4 and 5 antennas are even slightly smaller than the interference using 2 antennas). The reason is because of GA, which cannot guarantee to find global optimal solution every time. However, even the GA cannot guarantee global optimal with 100 percent probability, the derived values are still close to our bound.

The antennas' optimal weights derived by GA are shown in Table. 1. We also show the interference power generated on the primary user to verify the correctness of our solution. From Fig. 7, we can see that the interferences generated on the primary user's receiver are all smaller than -130 dB, which has nearly no interference to the primary user.

6 Conclusion

In this paper, we addressed the spectrum abuse problem. We proposed a spectrum access enforcing mechanism based on cooperative jamming. Our method doesn't rely on any modification of existing devices or protocols. We consider the uncertainty of channel conditions from guardian to abusers and the uncertainty of abusers' locations. We formulate our problem into an optimization problem and derive a lower bound of it. In the evaluation we use Genetic Algorithm to solve our problem and compare the derived maximum interference to abusers with the bound. We also analyzed the secure range provided by the guardian and the relation of generated interference to abuser with the number of antennas of guardian.

References

1. Atia, G., Sahai, A., Saligrama, V.: Spectrum enforcement and liability assignment in cognitive radio systems. In: Proceedings of the 3rd IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, Chicago IL (2008)
2. Bendov, O., Browne, J., Rhodes, C., Wu, Y., Bouchard, P.: Dtv coverage and service prediction, measurement and performance indices. IEEE Transactions on Broadcasting 47(3), 207–217 (2001)
3. Dong, L., Han, Z., Petropulu, A., Poor, H.: Cooperative jamming for wireless physical layer security. In: IEEE/SP 15th Workshop on Statistical Signal Processing, SSP 2009, pp. 417–420 (2009)
4. Dong, L., Han, Z., Petropulu, A., Poor, H.: Improving wireless physical layer security via cooperating relays. IEEE Transactions on Signal Processing 58(3), 1875–1888 (2010)
5. Gollakota, S., Adib, F., Katabi, D., Seshan, S.: Clearing the rf smog: making 802.11n robust to cross-technology interference. In: Proceedings of the ACM SIGCOMM 2011 Conference, SIGCOMM 2011, pp. 170–181. ACM, New York (2011)
6. Haenggi, M.: A geometric interpretation of fading in wireless networks: Theory and applications. IEEE Transactions on Information Theory 54(12), 5500–5510 (2008)
7. Jia, J., Zhang, Q., Zhang, Q., Liu, M.: Revenue generation for truthful spectrum auction in dynamic spectrum access. In: Proceedings of the Tenth ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc 2009, pp. 3–12. ACM, New York (2009)

8. Kasbekar, G.S., Sarkar, S.: Spectrum auction framework for access allocation in cognitive radio networks. In: Proceedings of the Tenth ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc 2009, pp. 13–22. ACM, New York (2009)
9. Kelif, J.M., Coupechoux, M.: Joint impact of pathloss shadowing and fast fading - an outage formula for wireless networks. CoRR, abs/1001.1110 (2010)
10. Liu, S., Greenstein, L.J., Trappe, W., Chen, Y.: Detecting anomalous spectrum usage in dynamic spectrum access networks. Ad Hoc Networks 10(5), 831–844 (2012); Special Issue on Cognitive Radio Ad Hoc Networks
11. Liu, Y., Ning, P., Dai, H.: Authenticating primary users' signals in cognitive radio networks via integrated cryptographic and wireless link signatures. In: 2010 IEEE Symposium on Security and Privacy (SP), pp. 286–301 (2010)
12. Negi, R., Goel, S.: Secret communication using artificial noise. In: 2005 IEEE 62nd Vehicular Technology Conference, VTC-2005-Fall, vol. 3, pp. 1906–1910 (2005)
13. Proakis, J., Salehi, M.: Digital communications. McGraw-Hill Higher Education (2008)
14. Rappaport, T.S.: Wireless Communications: Principles and Practice, 1st edn. IEEE Press, Piscataway (1996)
15. Sahai, A., Woyach, K.A., Atia, G., Saligrama, V.: A technical framework for light-handed regulation of cognitive radios. Comm. Mag. 47(3), 96–102 (2009)
16. Vo-Huu, T.D., Blass, E.-O., Noubir, G.: Counter-jamming using mixed mechanical and software interference cancellation. In: Proceedings of the Sixth ACM Conference on Security and Privacy in Wireless and Mobile Networks, WiSec 2013, pp. 31–42. ACM, New York (2013)
17. Shen, W., Ning, P., He, X., Dai, H.: Ally friendly jamming: How to jam your enemy and maintain your own wireless connectivity at the same time. In: Proceedings of the 2013 IEEE Symposium on Security and Privacy, SP 2013. IEEE Computer Society (2013)
18. Li, B., Feng, X., Zhang, Q.: Enabling co-channel coexistence of 802.22 and 802.11af systems in tv white spaces. In: 2013 IEEE International Conference on Communications (ICC) (2013)
19. Yang, L., Zhang, Z., Zhao, B.Y., Kruegel, C., Zheng, H.: Enforcing dynamic spectrum access with spectrum permits. In: Proceedings of the Thirteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc 2012, pp. 195–204. ACM, New York (2012)

Photo Forensics on Shanzhai Mobile Phone

Gang Zhou[1], Yanbin Tang[2,⋆], Junbin Fang[2,3,⋆⋆], Zoe L. Jiang[4], K.P. Chow[2],
S.M. Yiu[2], Lucas C.K. Hui[2], Rougsheng Xu[5], Yonghao Mai[6],
Shuhui Hou[7], and Fei Xu[8]

[1] Wuhan Engineering Science and Technology Institute, Wuhan, China
garretzhou@163.com

[2] Department of Computer Science, The University of Hong Kong, HKSAR, China
{chow,ybtang,jbfang,smyiu,hui}@cs.hku.hk

[3] Department of Optoelectronic Engineering, Jinan University, Guangzhou, China
junbinfang@gmail.com

[4] Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
zoeljiang@gmail.com

[5] Institute of High Energy Physics, Chinese Academy of Sciences, China
xurs@ihep.ac.cn

[6] Hubei University of Police, China
myh9999@163.com

[7] Department of Information and Computing Science,
University of Science and Technology Beijing, China
shuhui@ustb.edu.cn

[8] Institute of Information Engineering, Chinese Academy of Sciences, China
xufei@iie.ac.cn

Abstract. There is an increasing number of crime cases involving mobile phones. In particular, due to the low cost of China shanzhai mobile phone (Chinese pirated mobile phone), a significant portion of these crime cases (all over the world) is related to these shanzhai phones. Quite a number of the cases also involve pictures. The difficulty of conducting forensic investigation on shanzhai phones is the lack of specifications. In this paper, we try to provide some important information of how a photo is stored inside a MTK-based shanzhai phone (one of the most popular platforms for shanzhai phones), and provide a method to recover deleted photos from the physical segments of flash memory of a shanzhai phone. *abstract* environment.

Keywords: Shanzhai mobile phone, photo forensics, file carving, MediaTek.

1 Introduction

In the last decade, worldwide mobile phone usage has increased dramatically. Globally, The number of mobile cellular subscriptions reached 5.3 billion by the

⋆ Joint first authors.

⋆⋆ Corresponding author.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 454–464, 2013.

end of 2010, reported by the International Telecommunication Union (ITU). And vendors shipped 371.8 million units in Q1 2011, growing 19.8 percent year-over-year (IDC) [1]. At the same time, the computational power and storage of mobile phones are getting more and more powerful, especially the deployment of dual-core CPU and gigabytes of internal memory [2]. Due to the mobility and the portability, mobile phones have almost become people's belongings, and quite often be involved in some criminal cases [3]. More seriously, the powerful mobile phone can be used as a criminal tool anytime and anywhere. In both cases, a lot of digital evidences may stored inside the mobile phone and mobile phone forensics techniques are necessary for retrieving and investigating the information.

Mobile phone forensics has been studied for quite a long period and there are several commercial products for investigating the mobile phones of world leading brands, such as Symbian mobile phone, Android mobile phone, Blackberry mobile phone, iPhone and etc. However, in China, a new category of mobile phone with a commonly known brand of "Shanzhai mobile phone" ('Shanzhai phone for short) emerged from 2007 after China's government removed the license policy to manufacture mobile phones and now it is flooding in the global mobile phone market due to its high cost-performance ratios [4]. The Chinese word "Shanzhai" originally means "mountain village", but now it has another meaning to refer to imitation, low-end and unprofessional brands and goods, particularly electronics. Contrast to the remarkable growth of "Shanzhai phone", there is little published research work related to Shanzhai phone forensics. The reason may lie in the shortage of the technical documents of Shanzhai phone and the great number of Shanzhai phone models. Benefit from the turn-key solution provided by MediaTek (MTK) [5] and Spreadtrum [6], the development period for Shanzhai phone can be shortened from over 1 year to 1 month. It means that there will be thousands of models of Shanzhai phone appeared in market during one year. Unfortunately, such a quick changing becomes a nightmare to researchers to perform digital forensics on Shanzhai phone. But, since Shanzhai phone is spreading worldwide and there is an increasing trend that Shanzhai phones are found to be used by criminals in many crime cases, getting deeper investigation on Shanzhai phone is necessary and Shanzhai phone forensics unavoidably becomes more and more important.

Nowadays, mobile phone is usually equipped with one in-built camera, sometimes two cameras, to provide more enhanced multimedia functions. And the resolution the in-built cameras is getting higher, typically 5 to 8 Mega pixels at present. This enhancement increased the quality of the photo or video taken by the in-built camera, and then mobile phone is used more frequently in user's daily activity, even in working. For example, more and more people use mobile phone's in-built camera to take screenshot of slides during a presentation instead of using a specific camera as that means he needs to carry one more digital device. Therefore, when a mobile phone is involved in a digital crime case, it is highly possible that there are vast multimedia files inside the mobile phone, which may be relevant to the crime case. In this paper, we attempt to provide some important information of how the photo are stored inside a MTK-based Shanzhai phone (the most popular platform for Shanzhai phones), and then study how to recover deleted photo file from the physical segments of flash memory of Shanzhai phone. With JPEG file signature identification techniques, we investigate how to reconstruct JPEG file from sequentially distributed fragments and bi-fragmented data segments. This represents the first step towards photo forensics on Shanzhai mobile phone.

2 Related Works

There has been some research on mobile phone forensics since early 2000s. There are a wide range of mobile forensic tools developed to acquire data from the flash memory of mobile phones [7]. However, most of the tools use commands and response protocols to indirectly access the memory. These commands and protocols depend on the operating system and actually change the contents of the memory. Only data visible to the operating system can be recovered. Also, such tools fail to retrieve data from dead or faulty mobile phones. Another problem with such tools is that they cannot recover deleted data. Flasher tools are the easiest and non-invasive way to read flash memory data [8], which have been used in quite a few mobile forensic cases [9,10]. Both the above approaches cannot ensure a complete dump of the memory and may depend a lot on the operating system. Also, if the data connector of the mobile phone is not supported by the flasher tool or the pins of the data connector is not connected directly to the communication pins of the main processor, electronic work of wiring the communication pins out from the Printed Circuit Borad (PCB) to connect the flasher tool may be required. Physical extraction approach is to physically remove the internal flash memory chip from the mobile phone and read it with a memory reader. This procedure requires professional engineers since memory chips may be damaged during de-soldering. Joint Test Action Group (JTAG) is an embedded test technique to test automatically the functionality and quality of the soldered integrated components on Printed Circuit Board (PCB), which is a standard test access port and boundary-scan architecture. It controls the phone's micro-processor in debug mode to communicate with the memory chip, and dump the memory bit-by-bit. Therefore, it ensures the completeness of the forensics binary image, and is operating system independent. There has been some research work on mobile forensics using JTAG [11,12].

3 Physical Data Storage and Logical File System in Shanzhai Phone

In this paper, a typical model of Shanzhai phone is selected to be studied in the experiments. The model is an imitated version of Apple iPhone 4, as shown in Figure 1. This model is based on one low-end processor of MediaTek, MT6253, which

Fig. 1. A typical Shanzhai phone (imitated version of Apple iPhone 4)

is Mediatek's first monolithic GSM/GPRS handset chip solution which offers highest level of integration with lowest power consumption and best-in-class features. Such that most of Shanzhai phone models were developed on this platform.

Inside the Shanzhai phone, a 16M bytes NOR flash chip (Toshiba TC58FYM7T8C) integrated with a 4M bytes RAM is used to work as read-only memory (ROM) for operating system code and as non-volatile random access memory (NVRAM) for data storage. As shown in Figure 2, the 16M bytes NOR flash of the Shanzhai phone is divided into two parts. The first 14M bytes (memory address from 0–0xE00000) are used to store code and will be kept unchanged after the Shanzhai phone is produced. Noted that this is the default configuration in MTK development solution.

The remaining 2M bytes (memory address from 0xE00000–0xFFFFFF) are further divided logically into two areas. As shown in Figure 3, both of the two areas can be seen as removable drive under Windows OS when the Shanzhai phone is connected to a computer via USB data cable. Note that this is only the logical distribution of the two areas. Physically, the blocks in flash for the two areas are mixed and not separated as clear as this figure. From the partition information, we know that both drives are formatted in FAT12 format, but only the drive (here is drive H:) corresponding to USER area can be accessed via Windows, the other one (drive I:) corresponding to SYSTEM area cannot be read, write or viewed by a normal user. In general, the USER area is kept for the Shanzhai phone user as a storage to exchange data between the phone and a computer, while the SYSTEM area is kept for the OS of phone as a virtual memory to save the data managed by OS. Note that some of the data saved in SYSTEM area can be viewed or edited by user via the user interface (UI) of the Shanzhai phone, such as the settings of the phone, phonebook, call log, SMS and etc.

With the help of a flasher tool, the total 16M bytes of raw data in the flash memory can be retrieved as a memory dump and can be further investigated in a computer as a binary file. The complete physical dump enable us to investigate the photo files on Shanzhai phone, including the deleted ones.

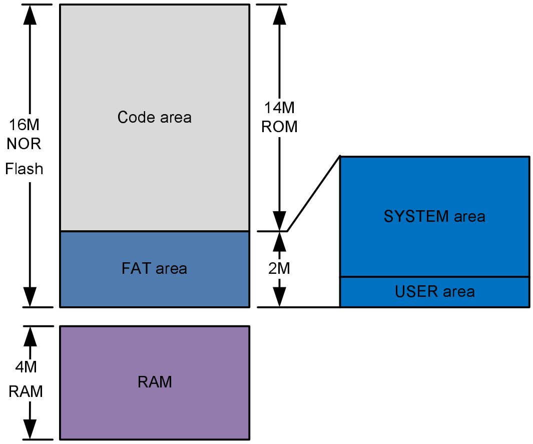


Fig. 2. The NOR flash memory for the Shanzhai phone

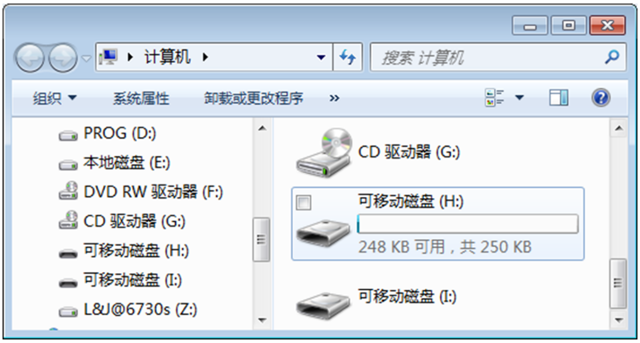


Fig. 3. The directory viewed from Windows

4 Photo Forensics in Shanzhai Phone

In the Shanzhai phone's memory dump, all the data is stored in binary format, regardless of what the type of contents it represents. To carry out forensics investigation on the photo of the Shanzhai phone, the main difficulty and the first and the foremost task is recovering (deleted) photo data from the binary dump. In this section, we focuses on JPEG files, the most famous type of digital photo, and the recovery technology of (deleted) JPEG files from Shanzhai phone's memory dump file will be explained in details. The recovery includes two stages. The first stage is to identify all the JPEG file fragments stored in the memory, typically, 512 bytes per fragment. Then the identified fragments will used for the second stage to reassemble the original full or partial photo.

In our previous research [13,14], we found that the photo files in Shanzhai phone are stored in a block-based reverse order on the flash memory. Photos captured by in-built camera and those copied from outside via PC have the same characteristic of storage allocation. Therefore, in the simplest case, after the identification of JPEG data fragments, just cascading all the data blocks in a reversed backward order can easily reconstruct the expected JPEG photo. To identify whether a data block in the flash memory is a fragment of JPEG file, some previous work about file type classification methods [15,16,17] could be applied to identify the file type of data fragments from extracted Shanzhai phone memory dump. Then the next problem is to try to find out methodologies to reassemble JPEG data fragments. Two schemes were applied to recover some photos stored on Shanzhai mobile phone. One is validating method to recover those sequential files. Another is bi-fragment gap carving (BGC) [19] technique to reassemble bi-fragmented files. Sequential file stores as contiguous location on storage media, whereas file fragmentation occurs when a single file has been broken into multiple pieces. Specifically, bi-fragmented file means file has been split into exactly two pieces. In Figure 4, file B locates on contiguous data blocks which is a sequential file, however file A keeps on two parts by means of bi-fragmented file.

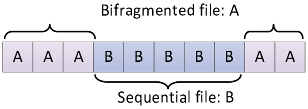


Fig. 4. Sequential file and bi-fragmented file

JPEG [18] apply lossy compression algorithm to gain high reputation of image quality with compression size. JPEG file start with specific signature 0xFFD8 as file header and end with 0xFFD9 as file footer. Moreover, the header of JPEG file holds a lot of information like the Huffman table, Quantization table, width, height etc., with specific predefined signatures for each segment. Table 1 gives a brief description of JPEG markers. As well as, each marker followed by some related parameters. Those data structures could provide important information to decode the compressed JPEG file and be used to identify the beginning and ending of JPEG file. In our experiment, quickly scan of reversed order dump, and identify those markers in specific locations. For start of image (SOI), it must appears at the beginning of one block as a validate signature to indicate a new file.

After JPEG file signature identification procession, if between one pair of file header and footer, all data blocks belong to same file type, a data object validator could be applied to indicate whether those data obey the data structure of JPEG. A JPEG decoder could be used as the object validator to check whether or not data block can be decoded successfully in sequence. If decoding successfully, those data will be extracted as a complete JPEG file and excluded from unrecovered data source. For fragmented JPEG files, especially bi-fragmented files, which means file locates on two different parts. BGC could be applied

Table 1. JPEG signatures [18]

Short name	Bytes	Name
SOI	0xFF, 0xD8	Start Of Image
SOF0	0xFF, 0xC0	Start Of Frame (Baseline DCT)
DHT	0xFF, 0xC4	Define Huffman Table(s)
DQT	0xFF, 0xDB	Define Quantization Table(s)
DRI	0xFF, 0xDD	Define Restart Interval
SOS	0xFF, 0xDA	Start Of Scan
RSTn	0xFF, 0xDn(n=0–7)	Restart
APPn	0xFF, 0xEn	Application-specific
EOI	0xFF, 0xD9	End Of Image

and it is very useful in small gap files. The algorithm of BGC is searching all combinations data blocks between JPEG file header and footer. At the mean time, excluding different number of data blocks until a successful decoding. If the JPEG file enable restart interval (RSTn), which increase in sequence from 0xFFD0, 0xFFD1, , to 0xFFD7 repeatedly with defined entropy-coded segment number, could be applied to identify the very next data block from other available candidate data blocks. When the fragmented files first part stop at 0xFFDi marker segment, then next block must contain restart marker $n = (i + 1) mod 8$. Figure 5 shows the whole process of file reassembly.

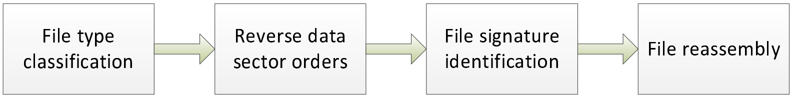


Fig. 5. The process of JPEG file reassembly

4.1 Experiments

Generally, the photos stored in mobile phone have two kinds of data source, including the photos captured by in-built camera and the photo files copied from PC. For Shanzhai phone, both kinds of photo data are basically stored in a same manner, i.e., in a block-based reverse order on the flash memory. Therefore, in our experiments, we focus more on the space distribution, especially the contiguity, of the data blocks storing photo fragments.

Four categories of operations were designed to simulate the scenarios in real life and to create different block distribution pattern of photo fragments. In our tests, the operations were performed as follows:

- (a) The simplest operations of creating photo data in Shanzhai phone, i.e., copying image files or taking photos, which most likely causes a sequentially distribution of data blocks.

- (b) Deleting image after creating it, which causes the loss of file information in filesystem, while the physical data is kept intact.
- (c) Copying big photo files from outside into Shanzhai phone, which will causes the bigger photo data being stored as bi-fragmented file.
- (d) Creating two photos, Image1 and Image2, on mobile phone. And then delete Image1, which is followed by creating a new photo, named Image3. The deleted photo may occasionally be overwritten partially due to the space recycle mechanism of flash memory controller.

To simplify the problem, in this paper, we assume that all the data blocks of JPEG file fragments can be identified correctly using the file type identification techniques. Therefore, our main targets include the recovery of the normal photos and deleted photos, both of which have sequentially allocated data blocks, and the recovery of the fragmented photos, for which the distribution of data blocks are separated into two areas with a gap between those two parts.

Figure 6 demonstrates the simplest case for photo recovery from Shanzhai phone memory dump, which corresponds to operations (a) and (b). The JPEG file is divided into 4 fragments. The first fragment, i.e., the header of the JPEG file, is stored in the data block at the largest physical address, marked as $i + 3$ in this sample. And the second fragment of the JPEG file is stored physically at the upper block, marked as $i + 2$. Following this pattern, the third fragment is stored at the $i + 1$ block and the footer fragment of the JPEG file is stored at the i block. Then after identifying the 4 fragments in the extracted physical memory dump, the recovery is quite straightforward. Reversing the order of the data blocks and then the blocks can be cascaded sequentially and be mapped to specific area of a JPEG structure, such that the JPEG file will be reconstructed for further investigation.

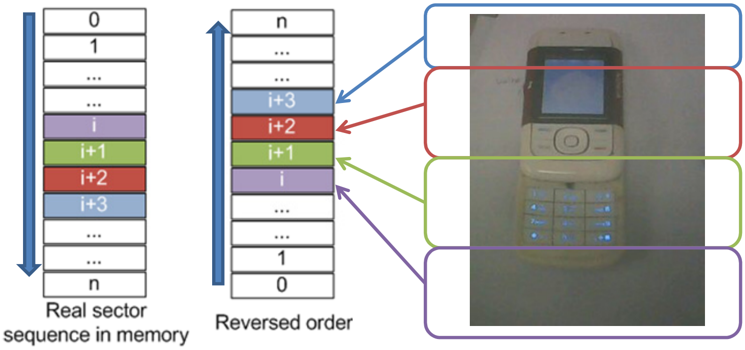


Fig. 6. Recovering photo from Shanzhai phone memory dump

For the photo files which is stored separately, the recovery work is more complex. Taking Figure 7 as an example, suppose there were more than one JPEG image files in the cell phone memory dump, whereas one picture, which named as Image1 , was fragmented into two parts, and JPEG data blocks from other picture file(s), shown as 1st part of Image1, 2nd part of Image 1, and data blocks j+1 and j from other JPEG files. To begin with, all fragmented data blocks of JPEG files are marked as gray after file type classification. As well as, during file signature identification procession, JPEG file signatures could be located, such as JPEG file header signature (0xFFD8) in block k+2, footer signature (0xFFD9) in block i and so on. In order to reassembly of bi-fragmented JPEG file, firstly, collecting all JPEG data blocks from sector 0 to sector n , which also means the sets of data blocks k+2, k+1, k, j+1, j, i+2, i+1, and i. Then, apply JPEG decoder as validator to verify data block combinations begin from JPEG header (block +2) to footer (block i), until a set of data blocks could be decoding successfully. If some errors happen during validating, current error data block will be removed and try the next data block. For example, the verification of JPEG will be started from data block k+2, and then concatenate with next data block k+1. If data block k+2 and k+1 could be decoded with no error, those two successfully decoded data blocks should be merged together. Contrary, when data blocks from different image file, some errors would happen during object validation. Such as errors may happen when data block j+1 concatenates after 1st part of Image1, then data block j+1 should be discarded from Image1 and keep on checking next data block until a successfully decoding of JPEG file.

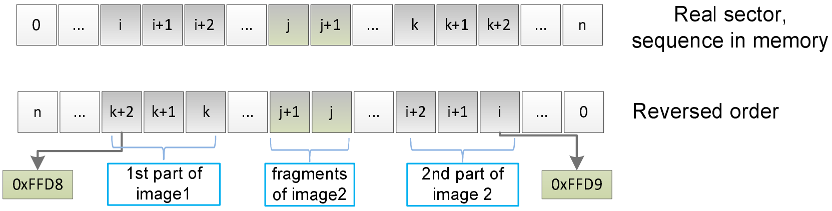


Fig. 7. Recovering bi-fragmented photo from Shanzhai phone memory dump

We carried out a series experiments for the operations (a)–(d) and the results are shown in Table 2. We performed 10 experiments for operations (a), 4 experiments for operations (b) and 3 experiments for operations (c) and (d). Using the specific recovery techniques of sequentially validating and BGC, all the JPEG files could be recovered successfully. Table 2 summarizes the details of the experiments.

Table 2. JPEG file recovery results

No. of test files	Operations	Recovery techniques	No. of recovered files
10	(a) Capturing/Copying	sequentially validating	10
4	(b) Deleting	sequentially validating	4
3	(c)&(d) Bi-fragmented	BGC	3

5 Conclusions

In this paper, we provide the first step towards photo forensics on China shanzhai mobile phones. In particular, we describe how a photo is stored inside a MTK-based shanzhai phone and provide a method to recover deleted photos from the segments of the flash memory of the phone. Future work includes extending the photo forensics to other types of shanzhai phones as well as considering forensics on other types of multimedia files such as video and audio files.

References

1. Worldwide Mobile Phone Market Grew 20İn: Q1 2011, Fueled By Smartphone Boom (2011), http://techcrunch.com/2011/04/28/worldwide-mobile-phone-market-grew-20-in-q1-fueled-by-smartphone-boom
2. Dual-core smartphones: The next mobile arms race (2011), http://www.silicon.com/technology/mobile/2011/01/12/dual-core-smartphones-the-next-mobile-arms-race-39746799
3. Mislan, R.: Cellphone crime solvers. IEEE Spectrum 47, 34–39 (2010)
4. Fake iPhone 4G Mobile Phone Hits Shanzhai Market (2010), http://www.suite101.com/news/fake-iphone-4g-mobile-phone-hits-shanzhai-market-a234058
5. MediaTek, http://www.mediatek.com
6. Spreadtrum, http://www.Spreadtrum.com
7. McCarthy, P.: Forensic analysis of mobile phones. Master thesis, University of South Australia (2005)
8. Breeuwsma, M., Jongh, M., Klaver, C., Knijff, R., Roeloffs, M.: Forensic Data Recovery from Flash Memory. Small Scale Digital Device Forensics Journal 1, 1–17 (2007)
9. Expert: 'Flasher' technology digs deeper for digital evidence, http://www.physorg.com/news95611284.html
10. Gratzer, V., Naccache, D.: Cryptography, Law Enforcement, and Mobile Communications. IEEE Security and Privacy 4, 67–70 (2006)
11. Willassen, S.: Forensic Analysis of Mobile Phone Internal Memory. In: IFIP The International Federation for Information Processing, Advances in Digital Forensics IV, pp. 191–204. Springer, Boston (2005)
12. Zhang, Z.W.: The research of MTK mobile phones flash file system recovery. Net-info Security 11, 34–36 (2010)
13. Fang, J., Jiang, Z.L., He, M., Yiu, S.M., Hui, L.C.K., Chow, K.P., Zhou, G.: Investigating and Analyzing the Web-based Contents on Chinese Shanzhai Mobile Phones. In: Seventh IEEE International Workshop on Systematic Approaches to Digital Forensic Engineering (IEEE/SADFE 2012), Vancouver BC, Canada, pp. 1–14 (2012)
14. Fang, J., Jiang, Z.L., Chow, K.P., Yiu, S.M., Hui, L.C.K., Zhou, G., He, M., Tang, Y.: Forensic Analysis of Pirated Chinese Shanzhai Mobile Phones. In: Eighth Annual IFIP WG 11.9 International Conference on Digital Forensics, Advances in Digital Forensics VIII, pp. 117–130. Springer, Boston (2012)
15. Karresand, M., Shahmehri, N.: Oscar-file type identification of binary data in disk clusters and RAM pages. In: IFIP International Federation for Information Processing, Security and Privacy in Dynamic Environments, vol. 201, pp. 413–424. Springer, Boston (2006)

16. Karresand, M., Shahmehri, N.: File type identification of data fragments by their binary structure. In: 2006 IEEE Workshop on Information Assurance, pp. 140–147. IEEE Press, New York (2006)
17. Calhoun, W.C., Coles, D.: Predicting the types of file fragments. Digital Investigation 5, s14–s20 (2008)
18. JPEG Standard (JPEG ISO/IEC 10918-1 ITU-T Recommendation T.81), `http://www.w3.org/Graphics/JPEG/itu-t81.pdf`
19. Garfinkel, S.: Carving Contiguous and Fragmented Files with Fast Object Validation. Digital Investigation 4S, s2–s12 (2007)

Local Information Storage Protocol for Urban Vehicular Networks

Bo Xie[1], Yingwen Chen[1], Ming Xu[1], and Yuangang Wang[2]

[1] Dept. of Network Engineering, School of Computer,
National University of Defense Technology,Changsha, China
xiebo@nudt.edu.cn, csywchen@gmail.com, xuming@nudt.edu.cn
[2] Huawei Technologies Co., Ltd., China
wangyuangang@huawei.com

Abstract. Local information dissemination is a potential application of VANETs. By cooperative communication among vehicles, the information could be stored in a certain size of region for some time. However, it is difficult to determine the size for time requirement. Moreover, it is not suitable to use too large size for the consideration of communication overhead on the whole network. This paper focuses on a common traffic flow in urban road network, and proposes a distributed protocol for persistent local information storage (LISP). We firstly give an approximated size of the region under the condition of time requirement. We propose a novel mechanism of header responsibility to suppress the impact of broadcast storm and reduce storage consumption on vehicles. The simulation results show good performance in various scenarios. Our protocol could satisfy the time requirement with much lighter overhead of storage consumption.

1 Introduction

Vehicular ad hoc networks (VANETs) which include vehicles which are equipped with radio interfaces and able to communicate with an infrastructure have started to attract attention from many researchers in both industry and academia. The potential applications of VANETs include safety related applications such as cooperative forward collision warning system, traffic signal violation warning, lane change warning, information/entertainment applications for back seat passengers and content based information dissemination.

Content based information dissemination enjoys wide applicability in these types of networks, ranging from traffic information and warnings, to parking availability, fuel prices, road conditions, and advertisements [1]. For example, In Fig. 1, at a certain point called POI, road works are scheduled or an accident happens on that road, we aim at letting the information called EOI (Event of Interest) *stick* to some areas called ROI (Region of Interest). While 3G networks can be used to offer these services, these come with a cost, both at network and hardware level, and with limitations in granularity and coverage. On the other hand, this information has geographic validity and temporal validity. This locally relevant content may be of little concern to the rest of the world. Storing it in a small region may be of much use

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 465–473, 2013.

with less cost than in a wide network. Moreover, the information may only be valid for a certain amount of time, for example one hour. Our approach integrates infrastructure with ad hoc approaches to decrease costs and increase coverage.

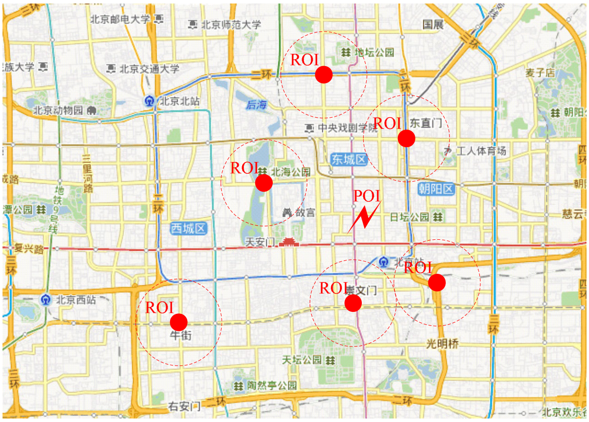


Fig. 1. Information Dissemination to Several ROI in Urban Scenarios

This approach has two steps. At first, the information of EOI is disseminated to these ROIs by some approaches, such as Internet or mobile vehicles, or infrastructures. The second step is to dissemination the EOI in this ROI.

This paper focuses on the second step. We propose a practical distributed protocol for persistent local information storage which purely depends on cooperative communication among vehicles. Once the EOI is disseminated to the vehicles in the ROI, the vehicles could 'store' the EOI by cooperative communication among them. The vehicles which receive the EOI will keep the copy for a certain time with the responsibility of keeping the EOI available for others. If, e.g., no (or too few) vehicles are around to replicate the EOI in the ROI, it will disappear (over time). The EOI may be tagged with a lifetime and are discarded thereafter. We introduce the performance metric of Mean Time to Information Loss (MTTIL) [2] to reflect the time that the EOI exists in the ROI. It is better of larger MTTIL. Both spatial-temporal requirements and storage consumption are our targets.

The remainder of this paper is organized as follows. Section 2 reviews the previous work, while Section 3 presents the motivation and protocol in details. Simulation and performance evaluations are presented in Section 4. Section 5 concludes this paper.

2 Related Work

There are several protocols of content dissemination for vehicular networks [1, 3-5]. Leontiadis et al. [1] describe a protocol for persistent content-based information dissemination in hybrid vehicular networks. There are three variants which are not independent but co-exist in this protocol to provide a single solution addressing content-based dissemination in heterogeneous environments.

The concept of floating content [6, 7] is similar to local information storage. They propose and analyze a fully distributed variant of an ephemeral content sharing service, solely dependent on the mobile devices in the vicinity using principles of opportunistic networking. Their analytical models show that such a system, without any supporting infrastructure, can be a viable and surprisingly reliable option for content sharing as long as a certain criterion, referred to as the criticality condition, is met.

Works targeting multicast communication in vehicular networks recently appeared in literature [8-14]. They used different versions of scoped epidemic protocols to constrain the propagation of a message within the given area specified by the publisher. Works in [15-17] define a notion of relevance to enable the routing layer to self-identify the areas in which the messages should be delivered.

Some other works in [18-20] focus on distributing content over an ad-hoc or DTN-like network, using the wireless network only as a kind of a cache for Internet content. They do not consider the case of managing content purely in the wireless domain.

The concept of time stable stored geocasting discussed in [21], is also similar to local information storage. In [22], the authors proposed a scheme for disseminating temporal information in highly partitioned mobile ad hoc network. While these works proposed some concepts or applications, they did not propose a practical distributed protocol for this local information storage application.

3 Dynamic Local Information Storage Protocol

3.1 Motivation and Requirements

As shown in Fig. 1, we focus on the phase of holding the EOI in the ROI by cooperative communication among vehicles. As the crossroad is quite common in urban scenarios, we study the problem in this situation.

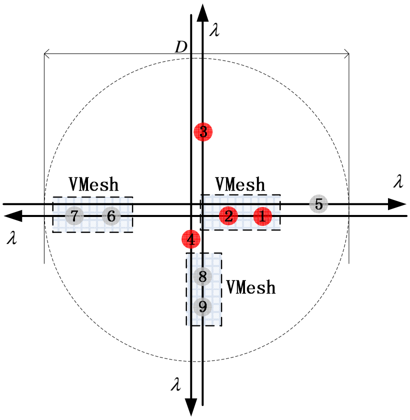


Fig. 2. An Illustrative Example for Crossroad Traffic

As illustrated by Fig. 2, a group of vehicles that are connected together forms a VMesh, in which each vehicle is within the transmission range of at least another vehicle. If one vehicle in a VMesh holds the information, it is assumed that all other vehicles in the same VMesh can receive the information through broadcast or multi-hop relay. For instance, node 1 and node 2 form a VMesh, node 6 and node 7 form another VMesh, and node 8 and node 9 form the other VMesh. The ROI has a length of *D*. When the EOI is disseminated in the ROI, node 1, 2, 3, 4 receive it. Then, they store the EOI, and move on. When node 1 meets node 5, which means that node 5 is in the transmission range of node 1, node 1 forwards the EOI to node 5. In the same way, node 5 will forward it to node 6, thereafter, the VMesh composed of node 6 and node 7 holds the EOI.

We assume the EOI has its expiration. Before the expiration, it is better that EOI exists longer and more vehicles receive the EOI. It is the spatial-temporal character of the EOI. The spatial and temporal requirements are both in the consideration. On the other hand, the spatial requirement determines the temporal property. For example, the EOI has an expiration of one hour. However, different places have different traffic flows. What is the size of the ROI for each place? However, we do not want to infinitely extend the size for the consideration of communication overhead on the network. Moreover, under the condition of time limitation, we want to reduce the storage consumption. In other words, if a vehicle has received the information and completed the mission of forwarding, it is better that it deletes the information from its cache as early as possible. In a word, the both spatial-temporal requirements and storage consumption are our targets.

3.2 Operation of Local Information Storage Protocol (LISP)

The message of EOI is identified by a unique message id. Besides the necessary bytes in the header of the message, an important triple <*Lifetime*, *D*, *Center*> should also be included in the header. *Lifetime* is the requested time the information existing for. For example, the information is available for more than one hour; therefore, *Lifetime* is 3600 seconds. *D* is the size of the region, and *Center* is the center point. With these two parameters, vehicles will know the boundary of the region, and could make decision for accepting, forwarding or deleting the information. However, *Lifetime* is closely related to *D*, in other words, the traffic around the center and *D* determine *Lifetime*. We have found the relationship among them, and propose an approximation model, as shown in Eq. (1). Therefore, we could firstly give an approximated *D* for a certain *Lifetime*. Once the EOI is disseminated, the maintenance for the EOI will completely depend on the local information storage protocol among vehicles.

$$MTTIL_{tw,cr} = \begin{cases} \dfrac{1}{C_4^2} e^{(2 \cdot D + 4 \cdot T_x)\lambda / v} \times MTTIL & D > T_x \\ \dfrac{1}{C_4^2} e^{4 \cdot D\lambda / v} \times \dfrac{D}{v} & D \le T_x \end{cases} \tag{1}$$

The protocol could be divided into two phases. The first phase is "store and forward". When a vehicle moves into the ROI, if it receives the information, it stores it and rebroadcasts it immediately. If it is the header of the VMesh, it takes the responsibility of forwarding the information to other VMesh. The header is defined as the vehicle which is the first one in the VMesh who has stored the information. As shown in Fig. 2, at this time, node 1 is the header of their VMesh. However, if some time later, node 1 deletes the information, node 2 becomes the header. When a header meets a new vehicle in the ROI, it will forward the information to the new vehicle. When a header receives the information from other vehicles, it will not rebroadcast it because it has stored it and the other members of the VMesh also have stored it. The mechanism of header responsibility suppresses the impact of broadcast storm, and reduces much storage consumption. If the vehicle is out of the ROI, and it hears the broadcast of other vehicles, it stores it but does not rebroadcast it. When it moves into the ROI, if it is the header, it rebroadcasts the information.

The second phase is "delete". There are two situations that vehicles will delete the information. 1) If a header is moving far away from the *Center*, and it is out of the ROI, it deletes the information. The header will send a notice to its neighbors; therefore, the second vehicle will know that it becomes the new header of the VMesh. 2) If the header has passed the *Center*, and has a distance of D_t away from the *Center*, and the number of its neighbors is larger than the threshold. Because the header takes the responsibility of forwarding the information to the vehicles on the other directions, specifically to the vehicles on the perpendicular roads, D_t should be shorter than the transmission range of vehicles T_x. The threshold could be computed by our approximation model. Because the vehicle length density ρ could be calculated as follows.

$$\rho = \frac{\lambda}{v} = \frac{N_{th}}{4 \cdot T_x} \tag{2}$$

In Eq. (2), N_{th} denotes the threshold number of neighbors around it. Therefore, according to Eq. (1) and Eq. (2), we could get the threshold.

4 Performance Evaluation

4.1 Simulation Setup

In this section, we conduct our simulation in Matlab which simplifies the wireless communication procedure ignoring the low-level issues, and strictly follows the assumptions in the analysis. We consider more on the up-level performance of the protocol than the communication level. We compare our LISP with the floating content protocol in [7] which is abbreviated to FCP in this paper. They use a constant size of ROI. When nodes are in the ROI, they always store the information, while they move out the side of the ROI, they delete it. They do not consider the neighbors density.

We assume all vehicles travel at a constant speed which is $v = 50km/h$, and the transmission range for all vehicles is $T_x = 50m$. As mentioned before, we assume the EOI needs to exist in the ROI for about one hour to two hours. Therefore, we firstly compute the range of ROI for different traffic scenarios. The parameters used in this simulation are listed as Table 1.

Table 1. Simulation Parameters

Arrival rate λ (/s)	**Range of ROI *D* (m)**
0.10	450
0.15	320
0.20	280
0.25	250

4.2 Simulation Results

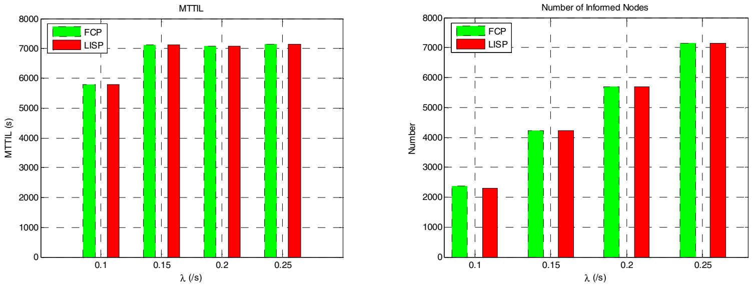
Our protocol LISP could reduce much storage consumption comparing with FCP in the same situation. For each scenario, we compare the MTTIL to show that our protocol could also guarantee the time limitation. We define the cost of storage consumption as follows.

$$Cost_i = TD_i - TR_i \tag{3}$$

$$CostTotal = \sum_{i}^{N} Cost_i \tag{4}$$

In Eq. (3), TD_i denotes the time at which vehicle *i* deletes the information, and TR_i denotes the time at which it receives the information. In Eq. (4), N means the total number of vehicles that have received the information. We also refer to these vehicles as informed vehicles. As mentioned before, it is better of larger N. Our protocol could decrease the average cost of each vehicle.

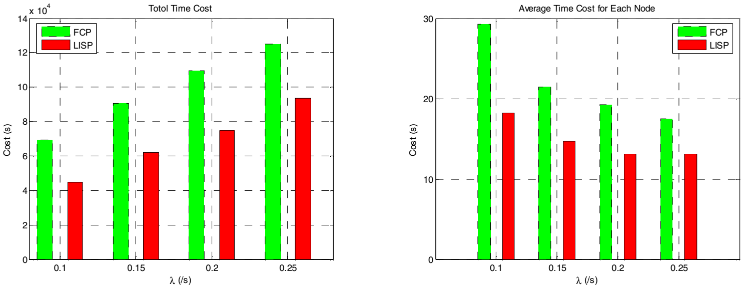
As show in Fig. 3(a), both of the two protocols could maintain the information about one hour to two hours, which satisfy the time requirement. MTTIL of our protocol LISP is almost the same as that of FCP. Fig. 3(b) also shows that the numbers of informed vehicles are almost the same. Both Fig. 3(a) and Fig. 3(b) show that LISP could satisfy the time requirement.



(a) (b)

Fig. 3. (a) MTTIL; (b) N, number of informed vehicles

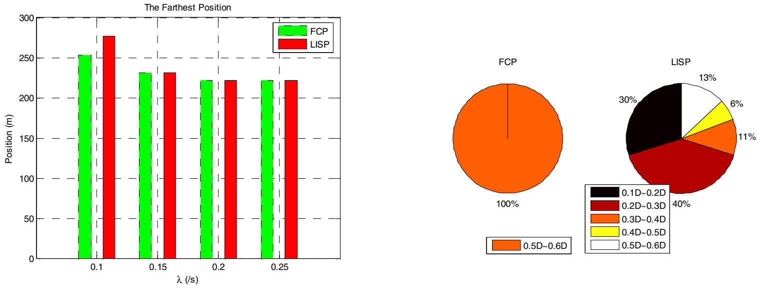
Our protocol is better than FCP for the reduction of storage consumption. As shown in Fig. 4(a) and Fig. 4(b), the total cost and average cost of LISP are both much lower than that of FCP. The reason is explained as follows. In FCP, vehicles always hold the information until they leave the ROI, therefore, the larger the ROI is, the higher average cost is, as shown in Fig. 4(b). With the increase of λ, more vehicles pass through the ROI, therefore, the total cost becomes larger.



(a) (b)

Fig. 4. (a) Total Cost; (b) Average Cost of Each Vehicle

We study the spatial influence of the two protocols. As pointed out, we do not want to infinitely extend the size for the consideration of communication overhead on the network. In other words, the farthest position at which vehicles delete the information could reflect this influence. Fig. 5(a) shows the farthest position away from the *Center*. The farthest position of LISP is a bit farther than that in FCP. However, with the increase of λ, the difference becomes unremarkable. Further insight into the "delete" position, we give the distribution of the "delete" position of the average four scenarios, as shown in Fig. 5(b). In FCP, all vehicles delete the information at the boundary area of the ROI, while in LISP, even about 70% of vehicles could delete the information before 0.3*D*, and only a few of them will delete at the boundary area. It reduces much communication overhead on the network out of the ROI.



(a) (b)

Fig. 5. (a) The Farthest Position; (b) Distribution of "delete" Position

5 Conclusion

Vehicles on the road could form temporal VMesh by cooperative communication among them. The local spatial-temporal related information may be stored in VMesh and transmitted among VMesh by "store and forward" without any supporting infrastructures. This paper proposed a distributed protocol for persistent local information storage LISP for a common traffic flow in urban road networks. Under the condition of time requirement, we firstly gave an approximated size of the region. The size should be large enough but not too large for the consideration of communication overhead on the whole network. Then, we proposed a novel mechanism of header responsibility to suppress the impact of broadcast storm and reduce storage consumption on vehicles. The simulation results showed good performance in various scenarios. Our protocol could satisfy the time requirement with much lighter overhead of storage consumption.

Acknowledgment. This research is partially supported by the National Science Foundation of China under Grant No. 61070211, No. 61070201 and No. 61003304; and Hunan Provincial Natural Science Foundation of China under grants No. 09JJ4034. The authors would like to express great appreciation to Xiaofang Zhou, Professor of Computer Science at the University of Queensland, for his suggestion on floating database.

References

1. Leontiadis, I., Costa, P., Mascolo, C.: Persistent content-based information dissemination in hybrid vehicular networks. In: Proceedings of IEEE International Conference on Pervasive Computing and Communications (PerCom 2009), Mannheim, Germany, pp. 1–10 (March 2009)
2. Liu, B., Khorashadi, B., Ghosal, D., Chuah, C., Zhang, M.: Assessing the VANET's local information storage capability under different traffic mobility. In: Proceedings of IEEE INFOCOM, Mini, San Diego, California, pp. 1–5 (March 2010)
3. Caliskan, M., Graupner, D., Mauve, M.: Decentralized discovery of free parking places. In: Proceedings of VANET 2006, Los Angeles, California, USA, pp. 30–39 (September 2006)
4. Sormani, D., Turconi, G., Costa, P., Frey, D., Migliavacca, M., Mottola, L.: Towards Lightweight Information Dissemination in InterVehicular Networks. In: Proceedings of VANET 2006, Los Angeles, California, USA, pp. 20–29 (September 2006)
5. Lee, U., Magistretti, E., Gerla, M., Bellavista, P., Corradi, A.: Dissemination and Harvesting of Urban Data Using Vehicular Sensing Platforms. IEEE Transactions on Vehicular Technology 58(2), 882–901 (2009)
6. Kangasharju, J., Ott, J., Karkulahti, O.: Floating content: Information availability in urban environments. In: Proceedings of IEEE PerCom Workshops, Mannheim, Germany, pp. 804–808 (March-April 2010)

7. Ott, J., Hyytiä, E., Lassila, P., Vaegs, T., Kangasharju, J.: Floating Content: Information Sharing in Urban Areas. In: Proceedings of IEE Pervasive Computing and Communication (PerCom) Conference, Seattle, pp. 136–146 (March 2011)
8. Eichler, S., Schroth, C., Kosch, T., Strassberger, M.: Strategies for context-adaptive message dissemination in vehicular ad hoc networks. In: Proc. of V2VCOM 2006 (July 2006)
9. Sormani, D., Turconi, G., Costa, P., Frey, D., Migliavacca, M., Mottola, L.: Towards Lightweight Information Dissemination in InterVehicular Networks. In: Proc. of VANET 2006 (2006)
10. Korkmaz, G., Ekici, E., Ozguner, F., Ozguner, U.: Urban multi-hop broadcast protocol for inter-vehicle communication systems. In: Proc. of VANET 2004 (2004)
11. Xu, B., Ouksel, A., Wolfson, O.: Opportunistic resource exchange in inter-vehicle ad-hoc networks. In: Proceedings of the IEEE International Conference on Mobile Data Management, MDM 2004 (2004)
12. Wolfson, O., Xu, B.: Opportunistic dissemination of spatio-temporal resource information in mobile peer to peer networks. In: Proceedings of 15th International Workshop on Database and Expert Systems Applications, pp. 954–958 (August-September 2004)
13. Dornbush, S., Joshi, A.: StreetSmart Traffic: Discovering and Disseminating Automobile Congestion Using VANET's. In: Proceedings of the 65th Vehicular Technology Conference, Dublin, Ireland (April 2007)
14. Lakas, A., Shaqfa, M.: Geocache: Sharing and Exchanging Road Traffic Information Using Peer-to-Peer Vehicular Communication. In: Proceedings of the 73rd Vehicular Technology Conference (VTC Spring), pp. 1–7 (May 2011)
15. Adler, C., Eigner, R., Schroth, C., Strassberger, M.: Context-Adaptive Information Dissemination in VANETs Maximizing the Global Benefit. In: CSN (2006)
16. Caliskan, M., Graupner, D., Mauve, M.: Decentralized discovery of free parking places. In: Proc. of VANET 2006 (2006)
17. Kosch, T., Schwingenschlgl, C., Ai, L.: Information Dissemination in Multihop Inter-Vehicle Networks - Adapting the Ad-hoc On-demand Distance Vector Routing Protocol (AODV). In: The 5th Int. Conf. on Intelligent Transportation Systems (2002)
18. Lenders, V., May, M., Karlsson, G., Wacha, C.: Wireless ad hoc podcasting. ACM/SIGMOBILE Mobile Comp. and Comm. Rev. (2008)
19. Leontiadis, I., Mascolo, C.: Opportunistic Spatio-Temporal Dissemination System for Vehicular Networks. In: Proc. 1st Int. ACM MobiSys Workshop MobiOpp (2007)
20. Karlsson, G., Lenders, V., May, M.: Delay-tolerant broadcasting. IEEE Transactions on Broadcasting (2007)
21. Maihöfer, C., Leimüller, T., Schoch, E.: Abiding geocast: time-stable geocast for ad hoc networks. In: Proceedings of Second ACM International Workshop on Vehicular Ad Hoc Networks (VANET 2005), Cologne, Germany, pp. 20–29 (September 2005)
22. Lindemann, C., Waldhorst, O.: Effective dissemination of presence information in highly partitioned mobile ad hoc networks. In: Proceedings of 3rd Annual IEEE Communications Society on Sensor and Ad Hoc Communications and Networks (SECON 2006), Reston, VA, USA, pp. 236–245 (September 2006)

Maximum Independent Set of Links with Power Control

Chao Ma, Fahad Al-dhelaan, and Peng-Jun Wan

Illinois Institute of Technology, Chicago, IL 60616

Abstract. This paper addresses the joint selection and power assignment of a largest number of given links which can communicate successfully at the same time under the physical interference model. For this optimization problem, we present a constant-approximation algorithm with improved performance over existing approximation algorithms. In addition, both the algorithm design and analysis are applicable to arbitrary path-loss exponent and arbitrary dimension of the deployment space.

1 Introduction

Consider a multihop wireless network consisting of a set V of nodes with maximum transmission power P. The strength of a transmitted signal attenuates with a path loss factor $\eta r^{-\kappa}$, where r is the distance from the transmitter, κ is *path-loss exponent* (a constant between 2 and 5 depending on the wireless environment), and η is the *reference loss factor*. The signal quality perceived by a receiver is measured by the *signal to interference and noise ratio* (*SINR*), which is the quotient between the power of the wanted signal and the total power of unwanted signals (i.e., interferences) and the ambient noise ξ. In order to correctly interpret the wanted signal, the SINR must exceed certain threshold $\sigma > 1$ under the physical interference model. Thus, for a communication from a node u to a node v to be possible even without any interference, the transmission power of u should exceed

$$p_0(u,v) := \sigma\xi_0 \|uv\|^{\kappa},$$

where $\|uv\|$ is the Euclidean distance between u and v and $\xi_0 = \frac{\xi}{\eta}$. Thus, the largest possible set of communication links is

$$L = \left\{(u,v) \in V^2 : p_0(u,v) < P, u \neq v\right\}.$$

For each $a = (u,v) \in L$, we use $p_0(a)$ to denote $p_0(u,v)$, and use $\ell(a)$ to denote its length $\|uv\|$. In addition, the distance between the sender of a link $a \in L$ and the receiver of a link $b \in L$ is denoted by $\ell(a,b)$.

For any $I \subseteq L$, let $\mathcal{P}(I)$ denote the set of power assignments $p \in (0,P]^I$ satisfying that for each $a \in I$,

$$\frac{p(a)\,\ell(a)^{-\kappa}}{\sum_{b \in I \setminus \{a\}} p(b)\,\ell(b,a)^{-\kappa} + \xi_0} > \sigma.$$

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 474–485, 2013.

Under power control, a set $I \subseteq L$ is said to be *independent* if $\mathcal{P}(I) \neq \emptyset$. Clearly, all links in an independent set I must be node-disjoint. Given a subset A of L, the problem of finding a largest independent set I of L and a power assignment $p \in \mathcal{P}(I)$ is referred to as **Maximum Independent Set of Links with Power Control**. Variants of this problem has been studied in [1,2,3,4,6,7]. In this paper, we present an improved approximation algorithm for this problem with better performance.

The following standard terms and notations are adopted throughout this paper. Consider a non-empty set E. For any real-valued function f on E and any $S \subseteq E$, $f(S)$ represents $\sum_{e\in S} f(S)$. For any real-valued function f on $E \times E$, $a \in E$, and $S \subseteq E$, $f(a,S)$ represents $\sum_{b\in S} f(a,b)$, and $f(S,a)$ represents $\sum_{b\in S} f(b,a)$. Let $\prec$ be an ordering of E. For any pair of links $a, b \in E$, $a \prec b$ means that a appears before b in the ordering $\prec$; $a \preceq b$ represents that either $a \prec b$ or $a = b$. Consider any $a \in E$ and any $B \subseteq A$. $a \prec B$ (respectively, $a \preceq B$) means that $a \prec b$ (respectively, $a \preceq b$) for each $b \in B$. We use $B_{\prec a}$ (respectively, $B_{\preceq a}$) to denote the set of links $b \in B$ satisfying that $b \prec a$ (respectively, $b \preceq a$).

2 Symmetric Conflict Systems

A symmetric conflict system is specified by a pair (A, ρ) where A is a non-empty finite set and ρ is a symmetric "conflict" function from A^2 to $\mathbb{R}_+ \cup \{\infty\}$. Consider a parameter $\phi > 0$. A subset I of A is said to be a *ϕ-restricted independent set* (IS) of A with respect to ρ if $\rho(I \setminus \{a\}, a) < \phi$ for each $a \in I$. Let ρ_ϕ be the function on A^2 defined by

$$\rho_\phi(a,b) = \min\{\phi, \rho(a,b)\}.$$

for any a and b in A. Then, (A, ρ_ϕ) is also a symmetric conflict system. Clearly, any ϕ-restricted IS of A with respect to ρ is also a ϕ-restricted IS of A with respect to ρ_ϕ, and vice versa. The average conflict of A with respect to ρ_ϕ defined to be

$$\frac{\sum_{a\in A} \rho_\phi(a, I \setminus \{a\})}{|A|} = \frac{\sum_{a\in A} \sum\limits_{b\in A\setminus\{a\}} \rho_\phi(a,b)}{|A|}.$$

The following theorem was proved in [5].

Theorem 1. *Let φ be the average conflict of A with respect to ρ_ϕ. If $\phi \geq \varphi$, A contains a ϕ-restricted IS of size at least $\left(1 - \frac{\varphi}{2\phi}\right)|A|$; otherwise A contains a ϕ-restricted IS of size at least $\frac{\phi}{2\varphi}|A|$.*

Consider an ordering $\prec$ of A and a parameter $\varphi > 0$. A subset J of A is said to be a *φ-restricted inductively independent set* (IIS) in $\prec$ if $\rho(J_{\prec a}, a) < \varphi$ for each $a \in J$. A φ-restricted IIS J in $\prec$ can be computed greedily as follows: Initially, J is empty. For each $a \in A$ in the ordering $\prec$, it is added to J if $\rho(J, a) < \varphi$. Such J is referred to as the *greedy φ-restricted IIS* of A in $\prec$. Let $\succ$ denote the reverse of the ordering $\prec$. The following lower bound on $|J|$ was proved in [5].

Theorem 2. *Suppose that A is a ϕ-restricted IIS in $\succ$ for some $\phi > 0$. Then, $|J| \geq \frac{|A|}{1+\phi/\varphi}$.*

For any $\gamma > 0$, the *forward γ-shield number* of a subset B of A in $\prec$ is defined to be the smallest integer k satisfying that for any $B' \subseteq B$ and any $a \in A$ with $a \prec B'$, there exists a subset S of B' satisfying that $|S| \leq k$ and for each $b \in B' \setminus S$, $\rho(a,b) \leq \gamma\rho(S,b)$. The following lower bound on $|J|$ was proved in [5].

Theorem 3. *Suppose that O is a ϕ-restricted IIS of A in $\prec$ for some $\phi > 0$ and its forward $\frac{\varphi}{\phi}$-shield number in $\prec$ is μ. Then, $|J| \geq |O|/\mu$.*

In the remaining of this section, we introduce two symmetric conflict systems associated with the wireless network described in Section 1 that will be used later in this paper. Consider a set $A \subseteq L$ and a power assignment p on A satisfying that for each $a \in A$, $p_0(a) < p(a) \leq P$. For any two links a and b in A_p, the *relative interference* from a to b, denoted by $RI_p(a,b)$, is defined as follows: If a and b share a common node, then $RI_p(a,b) = \infty$; otherwise,

$$RI_p(a,b) = \sigma \frac{p(a)\,\ell(a,b)^{-\kappa}}{(p(b) - p_0(b))\,\ell(b)^{-\kappa}}.$$

The *relative interference* between two links a and b in A is defined to be

$$\overline{RI}_p(a,b) = RI_p(a,b) + RI_p(b,a).$$

It's easy to verify for any $I \subseteq A$, $p \in \mathcal{P}(I)$ if and only if $RI_p(I \setminus \{a\}, a) < 1$ for each $a \in I$. The symmetric conflict system $\left(A, \overline{RI}_p\right)$ is referred to as the *relative interference system* of (A,p). For any $\phi > 0$, and a ϕ-restricted IS of A w.r.t. $\overline{RI}_p$ is also referred to as a ϕ-restricted *strongly independent set* (SIS) of A under p.

For any pair of links a and b in L, denote

$$\begin{aligned} \varrho_1(a,b) &= \sigma\left(\frac{\min\{\ell(a),\ell(b)\}}{\min\{\ell(a,b),\ell(b,a)\}}\right)^{\kappa}, \\ \varrho_2(a,b) &= \sigma^2\left(\frac{\ell(a)\,\ell(b)}{\ell(a,b)\,\ell(b,a)}\right)^{\kappa}, \\ \varrho(a,b) &= \varrho_1(a,b) + \varrho_2(a,b). \end{aligned}$$

For any $A \subseteq L$, the symmetric conflict system (A, ϱ) is referred to as the *geometric interference system* of A.

3 Strongly Independent Sets

Suppose that A is a non-empty subset A of L and q is a power assignment on A satisfying that $p_0(a) < q(a) \leq P$. The power assignment p on A with $p(a) = P$ for each $a \in A$ is referred to as the canonical uniform power assignment on A. Let $\prec$ denote an ordering of A in the the *increasing* order of length. In this section, we establish the following properties of ϕ-restricted SIS's of A.

Theorem 4. *Suppose that I is a ϕ-restricted SIS of A under q where $\phi = \frac{1}{8\left(2+(8\sigma)^{-1/\kappa}\right)^{\kappa}}$, and p is the canonical uniform power assignment on A. If $p_0(a) \geq P/4$ for any $a \in I$, then I is also a 4ϕ-restricted SIS under p, and its forward 8ϕ-shield number in $\prec$ w.r.t. $\overline{RI}_p$ is at most 4.*

Theorem 5. *Suppose that I is a ϕ-restricted SIS of A under q for some $\phi > 0$. Then, I is also a ϕ'-restricted IS w.r.t. under ϱ where*

$$\phi' = \frac{\phi}{2}\left(\sqrt{1+\left(\frac{\phi}{2\sigma}\right)^{\frac{2}{\kappa}}}+\left(\frac{\phi}{2\sigma}\right)^{\frac{1}{\kappa}}\right)^{\kappa}+\frac{\phi^2}{4}.$$

In addition, if $\phi \leq 4^{-\kappa}$ then w.r.t. ϱ the forward $\frac{3}{2}\beta^{\kappa}$-shield number of I in $\prec$ is at most 5 where $\beta = \frac{2}{1-(\sigma/\phi)^{-1/\kappa}}$.

The proof of the above two theorems will utilize the following geometric fact discovered in [5]. Consider a positive parameter β. For a link $a \in A$ and a set B of disjoint links in A, a subset S of B is said to be a *β-guard* of B from a if for each link $b \in B \setminus S$,

$$\min_{a' \in S} \ell(b, b') \leq \beta \ell(b, a),$$
$$\min_{b' \in S} \ell(b', b) \leq \beta \ell(a, b).$$

The *β-guard number* of a set $B \subseteq A$ is defined to be the smallest integer k satisfying that for any $B' \subseteq B$ and any $a \in A$, there is a β-guard of B' from a whose size is at most k. The following lemma was proved in [5].

Lemma 1. *For any $\phi \in (0, \sigma)$, the β-guard number of any ϕ-restricted SIS of A under q is at most 4, where $\beta = \frac{2}{1-(\sigma/\phi)^{-1/\kappa}}$.*

We proceed to prove Theorem 4 and Theorem 5 in the next two subsections.

3.1 Proof of Theorem 4

For any pair of links a and b in I,

$$\frac{RI_p(a,b)}{RI_q(a,b)} = \frac{p(a)}{q(a)}\frac{q(b)-p_0(b)}{p(b)-p_0(b)} = \frac{P}{q(a)}\frac{q(b)-p_0(b)}{P-p_0(b)} \leq \frac{P}{q(a)} < 4,$$

and hence $RI_p(a,b) < 4RI_q(a,b)$. So, I is also a 4ϕ-restricted SIS under p. Consider any $I' \subseteq I$ and any $a \in A$ with $a \prec I'$. Let S be a minimum $(8\phi)^{-1/\kappa}$-guard of I from a. Since

$$\frac{2}{1-(\sigma/\phi)^{-1/\kappa}} = (8\phi)^{-1/\kappa},$$

$|S| \leq 4$ by Lemma 1. Consider any $b \in I' \setminus S$. Let b_1 and b_2 be the links in S satisfying that

$$\max\left\{\frac{\ell(b,b_1)}{\ell(b,a)}, \frac{\ell(b_2,b)}{\ell(a,b)}\right\} \leq (8\phi)^{-1/\kappa}.$$

Since $a \prec b_1$ and $a \prec b_2$, we have

$$\max\left\{\frac{RI_p(b,a)}{RI_p(b,b_1)}, \frac{RI_p(a,b)}{RI_p(b_2,b)}\right\} \leq 8\phi.$$

Thus,

$$\overline{RI}_p(a,b) \leq 8\phi\left(RI_p(b,b_1) + RI_p(b_2,b)\right) = 8\phi\overline{RI}_p(S,b).$$

So, the forward 8ϕ-shield number of J in $\prec$ w.r.t. $\overline{RI}_p$ is at most 4.

3.2 Proof of Theorem 5

The first part of Theorem 5 is a direct consequence of the lemma below.

Lemma 2. *For any two disjoint links a and b in A,*

$$\varrho_1(a,b) < \frac{\overline{RI}_q(a,b)}{2}\left(\sqrt{1+\left(\frac{\overline{RI}_q(a,b)}{2\sigma}\right)^{\frac{2}{\kappa}}} + \left(\frac{\overline{RI}_q(a,b)}{2\sigma}\right)^{\frac{1}{\kappa}}\right)^{\kappa},$$

$$\varrho_2(a,b) < \left(\frac{\overline{RI}_q(a,b)}{2}\right)^2.$$

Proof. Since

$$\left(\frac{\overline{RI}_q(a,b)}{2}\right)^2 > RI_q(a,b)\,RI_q(b,a) > \sigma\frac{q(a)}{q(b)}\left(\frac{\ell(b)}{\ell(a,b)}\right)^{\kappa}\sigma\frac{q(b)}{q(a)}\left(\frac{\ell(a)}{\ell(b,a)}\right)^{\kappa}$$
$$= \sigma^2\left(\frac{\ell(a)\,\ell(b)}{\ell(a,b)\,\ell(b,a)}\right)^{\kappa} = \varrho_2(a,b),$$

the second inequality holds.

Next, we prove the first inequality. Let

$$t = \frac{\ell(a,b)\,\ell(b,a)}{\ell(a)\,\ell(b)}.$$

Then,

$$t > \left(\frac{\overline{RI}_q(a,b)}{2\sigma}\right)^{-2/\kappa}.$$

We further claim that

$$\frac{\min\{\ell(a,b), \ell(b,a)\}}{\min\{\ell(a), \ell(b)\}} \geq \sqrt{1+t} - 1.$$

Assume to the contrary that the claim does not hold. We further assume by symmetry that $\ell(a,b) \leq \ell(b,a)$. Then,

$$\ell(a,b) < \left(\sqrt{1+t}-1\right)\min\{\ell(a),\ell(b)\}.$$

So,

$$\ell(b,a) \leq \ell(a)+\ell(a,b)+\ell(b) \leq \left(\sqrt{1+t}+1\right)\max\{\ell(a),\ell(b)\}.$$

Thus,

$$\begin{aligned}&\ell(a,b)\,\ell(b,a) < \left(\sqrt{1+t}-1\right)\left(\sqrt{1+t}+1\right)\ell(a)\,\ell(b)\\ &= t\ell(a)\,\ell(b) = \ell(a,b)\,\ell(b,a),\end{aligned}$$

which is a contradiction. So, the claim holds. The claim implies that

$$\begin{aligned}&\varrho_1(a,b) \leq \sigma\left(\sqrt{1+t}-1\right)^{-\kappa} < \sigma\left(\sqrt{1+\left(\frac{\overline{RI}_q(a,b)}{2\sigma}\right)^{-\frac{2}{\kappa}}}-1\right)^{-\kappa}\\ &= \frac{\overline{RI}_q(a,b)}{2}\left(\sqrt{1+\left(\frac{\overline{RI}_q(a,b)}{2\sigma}\right)^{\frac{2}{\kappa}}}+\left(\frac{\overline{RI}_q(a,b)}{2\sigma}\right)^{\frac{1}{\kappa}}\right)^{\kappa}.\end{aligned}$$

So, the first inequality holds.

Now consider any $a \in I$. By Lemma 2,

$$\begin{aligned}&\varrho_1(a,I\setminus\{a\}) < \sum_{b\in I\setminus\{a\}}\frac{\overline{RI}_q(a,b)}{2}\left(\sqrt{1+\left(\frac{\phi}{2\sigma}\right)^{\frac{2}{\kappa}}}+\left(\frac{\phi}{2\sigma}\right)^{\frac{1}{\kappa}}\right)^{\kappa}\\ &= \frac{1}{2}\left(\sqrt{1+\left(\frac{\phi}{2\sigma}\right)^{\frac{2}{\kappa}}}+\left(\frac{\phi}{2\sigma}\right)^{\frac{1}{\kappa}}\right)^{\kappa}\overline{RI}_q(a,I\setminus\{a\})\\ &\leq \frac{\phi}{2}\left(\sqrt{1+\left(\frac{\phi}{2\sigma}\right)^{\frac{2}{\kappa}}}+\left(\frac{\phi}{2\sigma}\right)^{\frac{1}{\kappa}}\right)^{\kappa},\end{aligned}$$

and

$$\begin{aligned}&\varrho_2(a,I\setminus\{a\}) < \sum_{b\in I\setminus\{a\}}\left(\frac{\overline{RI}_q(a,b)}{2}\right)^2 \leq \left(\sum_{b\in I\setminus\{a\}}\frac{\overline{RI}_q(a,b)}{2}\right)^2\\ &= \frac{\left(\overline{RI}_q(a,I\setminus\{a\})\right)^2}{4} \leq \frac{\phi^2}{4}.\end{aligned}$$

Thus,

$$\varrho\left(a, I \setminus \{a\}\right) = \varrho_1\left(a, I \setminus \{a\}\right) + \varrho_2\left(a, I \setminus \{a\}\right)$$
$$< \frac{\phi}{2}\left(\sqrt{1+\left(\frac{\phi}{2\sigma}\right)^{\frac{2}{\kappa}}} + \left(\frac{\phi}{2\sigma}\right)^{\frac{1}{\kappa}}\right)^{\kappa} + \frac{\phi^2}{4} = \phi'.$$

So, the first part of Theorem 5 holds.

We continue to prove the second part of Theorem 5. Consider any $I' \subseteq I$ and any $a \in A$ with $a \prec I'$. Let S_1 be a minimum β-guard of I' from a. By Lemma 1, $|S_1| \leq 4$. Let

$$I_1' = \left\{b \in I : \sigma\left(\frac{\ell(b)}{\max\{\ell(b,a), \ell(a,b)\}}\right)^{\kappa} \leq \frac{1}{2}\right\},$$
$$I_2' = \left\{b \in I : \sigma\left(\frac{\ell(b)}{\max\{\ell(b,a), \ell(a,b)\}}\right)^{\kappa} > \frac{1}{2}\right\}.$$

If I_2' is empty, let $S = S_1$; otherwise, choose b_2 to be a shortest link in I_2' and let $S = S_1 \cup \{b_2\}$. Consider any $b \in I' \setminus S$. We first prove that $\varrho_1(a,b) \leq \beta^{\kappa}\varrho_1(S_1,b)$ in two cases.

Case 1: $\ell(a,b) \leq \ell(b,a)$. Let b_1 be a link in S_1 such that $\ell(b,b_1) \leq \beta\ell(b,a)$. We claim that $\varrho_1(a,b) \leq \beta^{\kappa}\varrho_1(b_1,b)$. Indeed, since $\ell(a) \leq \ell(b_1)$, we have

$$\frac{\varrho_1(a,b)}{\varrho_1(b_1,b)} = \left(\frac{\min\{\ell(a), \ell(b)\}}{\min\{\ell(b_1), \ell(b)\}}\right)^{\kappa}\left(\frac{\min\{\ell(b,b_1), \ell(b_1,b)\}}{\min\{\ell(b,a), \ell(a,b)\}}\right)^{\kappa}$$
$$\leq \left(\frac{\min\{\ell(b,b_1), \ell(b_1,b)\}}{\ell(b,a)}\right)^{\kappa} \leq \left(\frac{\ell(b,b_1)}{\ell(b,a)}\right)^{\kappa} \leq \beta^{\kappa}.$$

Thus, the claim holds.

Case 2: $\ell(a,b) > \ell(b,a)$. Let b_1 be a link in S_1 such that $\ell(b_1,b) \leq \beta\ell(a,b)$. Similar to **Case 1**, we can show that $\varrho_1(a,b) \leq \beta^{\kappa}\varrho_1(b_1,b)$.

Next, we prove that $\varrho(a,b) \leq 2\beta^{\kappa}\varrho(S,b)$ in two cases.

Case 1: $b \in I_1'$. Then,

$$\varrho_2(a,b) = \sigma\left(\frac{\ell(b)}{\max\{\ell(b,a), \ell(a,b)\}}\right)^{\kappa}\varrho_1(a,b) \leq \frac{1}{2}\varrho_1(a,b).$$

and hence

$$\varrho(a,b) = \varrho_1(a,b) + \varrho_2(a,b) \leq \frac{3}{2}\varrho_1(a,b) \leq \frac{3}{2}\beta^{\kappa}\varrho(S_1,b) \leq \frac{3}{2}\beta^{\kappa}\varrho(S,b).$$

Case 2: $b \in I_2'$. We claim that $\varrho_2(a,b) \leq \beta^{\kappa}\varrho_2(b_2,b)$. Indeed, by Lemma 2,

$$\ell(b_2)\,\ell(b) \leq \left(\frac{\phi}{2\sigma}\right)^{2/\kappa}\ell(b_2,b)\,\ell(b,b_2).$$

Since

$$\begin{aligned} \ell(a) \leq \ell(b_2) &\leq \ell(b), \\ \max\{\ell(b_2,a), \ell(a,b_2)\} &\leq (2\sigma)^{1/\kappa}\ell(b_2), \\ \max\{\ell(b,a), \ell(a,b)\} &\leq (2\sigma)^{1/\kappa}\ell(b), \end{aligned}$$

by the triangular inequality we have

$$\begin{aligned} &\ell(b_2,b)\ell(b,b_2) \leq [\ell(b_2,a)+\ell(a)+\ell(a,b)][\ell(b,a)+\ell(a)+\ell(a,b_2)] \\ =&\ell(a,b)\ell(b,a)+\ell(a,b_2)\ell(b_2,a)+\ell(a,b_2)\ell(a,b)+\ell(b_2,a)\ell(b,a) \\ &+\ell(a)[\ell(b_2,a)+\ell(a,b_2)+\ell(a,b)+\ell(b,a)]+\ell(a)^2 \\ \leq&\ell(a,b)\ell(b,a)+\left(3(2\sigma)^{2/\kappa}+4(2\sigma)^{1/\kappa}+1\right)\ell(b_2)\ell(b) \\ \leq&\ell(a,b)\ell(b,a)+\left(3(2\sigma)^{2/\kappa}+4(2\sigma)^{1/\kappa}+1\right)\left(\frac{\phi}{2\sigma}\right)^{\frac{2}{\kappa}}\ell(b_2,b)\ell(b,b_2) \\ =&\ell(a,b)\ell(b,a)+\left(3+4(2\sigma)^{-1/\kappa}+(2\sigma)^{-2/\kappa}\right)\phi^{\frac{2}{\kappa}}\ell(b_2,b)\ell(b,b_2) \\ \leq&\ell(a,b)\ell(b,a)+8\phi^{\frac{2}{\kappa}}\ell(b_2,b)\ell(b,b_2) \\ =&\ell(a,b)\ell(b,a)+\frac{1}{2}\ell(b_2,b)\ell(b,b_2). \end{aligned}$$

So,

$$\frac{\ell(b_2,b)\ell(b,b_2)}{\ell(a,b)\ell(b,a)} \leq 2.$$

Therefore,

$$\begin{aligned} \frac{\varrho_2(a,b)}{\varrho_2(b_2,b)} &= \frac{\left(\frac{\ell(a)\ell(b)}{\ell(a,b)\ell(b,a)}\right)^\kappa}{\left(\frac{\ell(b_2)\ell(b)}{\ell(b_2,b)\ell(b,b_2)}\right)^\kappa} = \left(\frac{\ell(a)}{\ell(b_2)}\right)^\kappa \left(\frac{\ell(b_2,b)\ell(b,b_2)}{\ell(a,b)\ell(b,a)}\right)^\kappa \\ &\leq \left(\frac{\ell(b_2,b)\ell(b,b_2)}{\ell(a,b)\ell(b,a)}\right)^\kappa \leq 2^\kappa \leq \beta^\kappa. \end{aligned}$$

So, the claim holds. Consequently,

$$\begin{aligned} \varrho(a,b) = \varrho_1(a,b)+\varrho_2(a,b) &\leq \beta^\kappa\varrho_1(S_1,b)+\beta^\kappa\varrho_2(b_2,b) \\ &\leq \beta^\kappa\varrho_1(S,b)+\beta^\kappa\varrho_2(S,b) = \beta^\kappa\varrho(S,b). \end{aligned}$$

Therefore, the second part of Theorem 5 holds.

4 Canonical Iterative Power Assignment

Let $\prec$ denote an ordering of L in the the *increasing* order of length. Consider a set I of short links. The *canonical iterative power assignment* p to I defined in [6] is fully determined by the criteria that for each $a \in I$ the total interference from $I_{\succ a}$ plus the noise is *exactly* $\frac{1}{2\sigma}$ times its wanted signal strength. Specifically, let $a_1, a_2, \cdots, a_{|I|}$ be the sequence of links in I sorted in $\prec$. Then

$$p\left(a_{|I|}\right) = 2p_0\left(a_{|I|}\right);$$

and for each i from $|I| - 1$ down to 1,

$$p\left(a_i\right) = 2p_0\left(a_i\right) + 2\sigma\ell\left(a_i\right)^{\kappa} \sum_{j=i+1}^{|I|} \frac{p\left(a_j\right)}{\ell\left(a_j, a_i\right)^{\kappa}}.$$

In [6], a sufficient condition on I for $\max_{a \in I} p\left(a\right) \leq P$ and I being independent under p was provided. In this section, we provide a looser sufficient condition.

Let φ be the unique root of

$$\varphi\left(1 + \left(2 + \left(\frac{\varphi}{\sigma}\right)^{1/\kappa}\right)^{\kappa}\right) = 1/4.$$

It is easy to show that

$$\frac{1}{4\left(1 + \left(2 + \frac{(4(2^{-\kappa}+1)\sigma)^{-1/\kappa}}{2}\right)^{\kappa}\right)} < \varphi < \frac{1}{4\left(1 + 2^{\kappa}\right)}.$$

Theorem 6. *If I is a φ-restricted IIS in $\prec$ w.r.t. ϱ and a $\frac{1}{4}$-restricted IIS in $\succ$ w.r.t. ϱ_1, then* $\max_{a \in I} p\left(a\right) \leq P$ *and I is independent under p.*

Proof. The proof is similar to the proof of Theorem 3 in [6]. We only highlight the sequence of the arguments and the key differences. The proof of $\max_{a \in I} p\left(a\right) \leq P$ is exactly the same as that in [6]. The independence of I under p is proved by showing that for each link $a \in I$ the total interference from $I_{\prec a}$ to a is less than $\frac{1}{2\sigma}$ times its wanted signal strength. In other words, for any $1 \leq i \leq |I|$,

$$\sum_{j=1}^{i-1} \frac{p\left(a_j\right)}{\ell\left(a_j, a_i\right)^{\kappa}} < \frac{1}{2\sigma} \frac{p\left(a_i\right)}{\ell\left(a_i\right)^{\kappa}}.$$

The above inequality holds trivially if $i = 1$. So, we assume that $i > 1$. It was shown in [6] that $\sum_{j=1}^{i-1} \frac{p(a_j)}{\ell(a_j, a_i)^{\kappa}}$ is equal to

$$\begin{aligned} & 2\sigma \sum_{k=2}^{i-1} \frac{p\left(a_k\right)}{\ell\left(a_k, a_i\right)^{\kappa}} \sum_{j=1}^{k-1} \left(\frac{\ell\left(a_k, a_i\right) \ell\left(a_j\right)}{\ell\left(a_k, a_j\right) \ell\left(a_j, a_i\right)}\right)^{\kappa} \\ & + 2\sigma \frac{p\left(a_i\right)}{\ell\left(a_i\right)^{\kappa}} \sum_{j=1}^{i-1} \left(\frac{\ell\left(a_i\right) \ell\left(a_j\right)}{\ell\left(a_i, a_j\right) \ell\left(a_j, a_i\right)}\right)^{\kappa} \\ & + 2\sigma \sum_{k=i+1}^{|I|} \frac{p\left(a_k\right)}{\ell\left(a_k, a_i\right)^{\kappa}} \sum_{j=1}^{i-1} \left(\frac{\ell\left(a_k, a_i\right) \ell\left(a_j\right)}{\ell\left(a_k, a_j\right) \ell\left(a_j, a_i\right)}\right)^{\kappa} + 2\xi_0 \sigma \sum_{j=1}^{i-1} \left(\frac{\ell\left(a_j\right)}{\ell\left(a_j, a_i\right)}\right)^{\kappa}. \end{aligned}$$

In addition, following the same argument as in [6] we can show that for any k between 2 and $|I|$ other than i,

$$\sigma \sum_{j=1}^{\min\{i,k\}-1} \left(\frac{\ell(a_j)\,\ell(a_k,a_i)}{\ell(a_j,a_i)\,\ell(a_k,a_j)} \right)^{\kappa} < \varphi \left(2 + \left(\frac{\varphi}{\sigma} \right)^{1/\kappa} \right)^{\kappa}.$$

Thus,

$$\begin{aligned}
&\sum_{j=1}^{i-1} \frac{p(a_j)}{\ell(a_j,a_i)^{\kappa}} \\
<&2\varphi \left(2 + \left(\frac{\varphi}{\sigma} \right)^{1/\kappa} \right)^{\kappa} \sum_{k=2}^{i-1} \frac{p(a_k)}{\ell(a_k,a_i)^{\kappa}} + \frac{2\varphi}{\sigma} \frac{p(a_i)}{\ell(a_i)^{\kappa}} \\
&+ 2\varphi \left(2 + \left(\frac{\varphi}{\sigma} \right)^{1/\kappa} \right)^{\kappa} \sum_{k=i+1}^{|I|} \frac{p(a_k)}{\ell(a_k,a_i)^{\kappa}} + 2\varphi\xi_0 \\
=&2\varphi \left(2 + \left(\frac{\varphi}{\sigma} \right)^{1/\kappa} \right)^{\kappa} \sum_{k=2}^{i-1} \frac{p(a_k)}{\ell(a_k,a_i)^{\kappa}} + \frac{2\varphi}{\sigma} \frac{p(a_i)}{\ell(a_i)^{\kappa}} \\
&+ 2\varphi \left(2 + \left(\frac{\varphi}{\sigma} \right)^{1/\kappa} \right)^{\kappa} \left(\frac{1}{2\sigma} \frac{p(a_i)}{\ell(a_i)^{\kappa}} - \xi_0 \right) + 2\varphi\xi_0 \\
=&2\varphi \left(2 + \left(\frac{\varphi}{\sigma} \right)^{1/\kappa} \right)^{\kappa} \sum_{k=2}^{i-1} \frac{p(a_k)}{\ell(a_k,a_i)^{\kappa}} + \frac{\varphi}{\sigma} \left(2 + \left(2 + \left(\frac{\varphi}{\sigma} \right)^{1/\kappa} \right)^{\kappa} \right) \frac{p(a_i)}{\ell(a_i)^{\kappa}} \\
&- 2\varphi \left(\left(2 + \left(\frac{\varphi}{\sigma} \right)^{1/\kappa} \right)^{\kappa} - 1 \right) \xi_0 \\
<&2\varphi \left(2 + \left(\frac{\varphi}{\sigma} \right)^{1/\kappa} \right)^{\kappa} \sum_{k=1}^{i-1} \frac{p(a_k)}{\ell(a_k,a_i)^{\kappa}} + \frac{\varphi}{\sigma} \left(2 + \left(2 + \left(\frac{\varphi}{\sigma} \right)^{1/\kappa} \right)^{\kappa} \right) \frac{p(a_i)}{\ell(a_i)^{\kappa}}.
\end{aligned}$$

Therefore,

$$\sum_{k=1}^{i-1} \frac{p(a_k)}{\ell(a_k,a_i)^{\kappa}} < \frac{\frac{\varphi}{\sigma} \left(2 + \left(2 + \left(\frac{\varphi}{\sigma} \right)^{1/\kappa} \right)^{\kappa} \right)}{1 - 2\varphi \left(2 + \left(\frac{\varphi}{\sigma} \right)^{1/\kappa} \right)^{\kappa}} \frac{p(a_i)}{\ell(a_i)^{\kappa}} = \frac{1}{2\sigma} \frac{p(a_i)}{\ell(a_i)^{\kappa}}.$$

This completes the proof of the theorem.

5 Power Control and Link Selection

In this section, we describe a simple algorithm **IS-PC** which computes an IS I of $A \subseteq L$ and a power assignment $p \in \mathcal{P}(I)$. The algorithm is described in Table 1. Let p_1 be the canonical uniform power assignment on A, and φ be the constant defined in the previous section. Denote

$$\begin{aligned}
A_1 &= \{a \in A : p_0(a) \geq P/4\}, \\
A_2 &= \{a \in A : p_0(a) < P/4\}.
\end{aligned}$$

The algorithm first computes a 1-restricted SIS I_1 of A under p_1, and then computes a subset I_2 of A_2 satisfying the conditions in Theorem 6 and the canonical iterative power assignment p_2 on I_2. If $|I_1| \geq |I_2|$, then (I_1, p_1) is the output solution; otherwise (I_2, p_2) is the output solution.

Table 1. The description of the aslgorithm **IS-PC**

```
Algorithm IS-PC:
$p_1 \leftarrow$ the canonical uniform power assignment on $A$;
$J_1 \leftarrow$ the greedy $\frac{1}{2}$-restricted IS of $A$ in $\prec$ w.r.t. $\overline{RI}_{p_1}$;
$I_1 \leftarrow$ the greedy $\frac{1}{2}$-restricted IS of $J_1$ in $\succ$ w.r.t $\overline{RI}_{p_1}$;
$A_2 \leftarrow \{a \in A : p_0(a) < P/4\}$;
$J_2 \leftarrow$ the greedy $\varphi$-restricted IS of $A_2$ in $\prec$ w.r.t. $\varrho$;
$I_2 \leftarrow$ the greedy $\frac{1}{4}$-restricted IS of $J_2$ in $\succ$ w.r.t. $\varrho_1$;
$p_1 \leftarrow$ the canonical iterative power assignment on $I_2$;
If $|I_1| \geq |I_2|$, return $(I_1, p_1)$; else return $(I_2, p_2)$.
```

Next, we derive an approximation bound of the algorithm **IS-PC**. Let

$$\phi_1 = \frac{1}{8\left(2 + (8\sigma)^{-1/\kappa}\right)^{\kappa}},$$

and ϕ_2 be the unique solution to the equation

$$\phi \frac{\left(\sqrt{1 + \left(\frac{\phi}{2\sigma}\right)^{2/\kappa}} + \left(\frac{\phi}{2\sigma}\right)^{1/\kappa}\right)^{\kappa} + \frac{\phi}{2}}{\left(1 - (\phi/\sigma)^{1/\kappa}\right)^{\kappa}} = \frac{4}{3}\frac{\varphi}{2^{\kappa}}.$$

Theorem 7. *The approximation ratio of the algorithm is at most* $\frac{32}{\phi_1} + \frac{20(1+4\varphi)}{\phi_2}$.

Proof. Let (O, q) be an optimal solution. We first show that $|O \cap A_1| \leq \frac{32}{\phi_1} |I_1|$. Let O_1 be a maximum ϕ_1-restricted SIS of $O \cap A_1$ under q. By Theorem 1, $|O_1| \geq \frac{\phi_1}{4} |O \cap A_1|$. By Theorem 4, with respect to $\overline{RI}_{p_1}$ the set O_1 is also a $4\phi_1$-restricted IS and its forward forward $8\phi_1$-shield number in $\prec$ is at most 4. By Theorem 3, $|J_1| \geq |O_1|/4$. By Theorem 2,

$$|I_1| \geq \frac{|J_1|}{2} \geq \frac{|O_1|}{8} \geq \frac{\phi_1}{32} |O \cap A_1|.$$

Next, we show that $|O \cap A_2| \leq \frac{20(1+4\varphi)}{\phi_2} |I_2|$. Let O_2 be a maximum ϕ_2-restricted SIS of $O \cap A_2$ under q. By Theorem 1, $|O_2| \geq \frac{\phi_2}{4} |O \cap A_2|$. Let

$$\gamma = \frac{3}{2}\left(\frac{2}{1 - (\phi_2/\sigma)^{1/\kappa}}\right)^{\kappa}.$$

Then,

$$\frac{\varphi}{\gamma} = \frac{\phi_2}{2}\left(\sqrt{1+\left(\frac{\phi_2}{2\sigma}\right)^{2/\kappa}} + \left(\frac{\phi_2}{2\sigma}\right)^{1/\kappa}\right)^{\kappa} + \frac{\phi_2^2}{4}.$$

Since $\phi_2 < \frac{4}{3}\frac{\varphi}{2^\kappa} < 4^{-\kappa}$, by Theorem 5, with respect to ϱ the set O_2 is also a $\frac{\varphi}{\gamma}$-restrict IS and its forward γ-shield number in $\prec$ is at most 5. By Theorem 3, $|J_2| \geq |O_2|/5$. By Theorem 2,

$$|I_2| \geq \frac{|J_2|}{1+4\varphi} \geq \frac{|O_2|}{5(1+4\varphi)} \geq \frac{\phi_2}{20(1+4\varphi)}|O \cap A_2|.$$

In summary,

$$\begin{aligned}|O| = |O \cap A_1| + |O \cap A_2| &\leq \frac{32}{\phi_1}|I_1| + \frac{20(1+4\varphi)}{\phi_2}|I_2| \\ &\leq \left(\frac{32}{\phi_1} + \frac{20(1+4\varphi)}{\phi_2}\right)\max\{|I_1|, |I_2|\}.\end{aligned}$$

So, the theorem holds.

Acknowledgements. This work was supported by the National Plan for Sciences & Technology (NPST) at King Saud University (Project No. 10-INF1184-02), the National Science Foundation of USA under grants CNS-0916666 and CNS-1219109, and by the National Natural Science Foundation of P. R. China under grant 61128005.

References

1. Fanghänel, A., Keßelheim, T., Vöcking, B.: Improved algorithms for latency minimization in wireless networks. In: Albers, S., Marchetti-Spaccamela, A., Matias, Y., Nikoletseas, S., Thomas, W. (eds.) ICALP 2009, Part II. LNCS, vol. 5556, pp. 447–458. Springer, Heidelberg (2009)
2. Halldórsson, M.M.: Wireless Scheduling with Power Control. In: Fiat, A., Sanders, P. (eds.) ESA 2009. LNCS, vol. 5757, pp. 361–372. Springer, Heidelberg (2009)
3. Kesselheim, T.: A Constant-Factor Approximation for Wireless Capacity Maximization with Power Control in the SINR Model. SIAM SODA, 1549–1559 (2011)
4. Kesselheim, T.: Approximation Algorithms for Wireless Link Scheduling with Flexible Data Rates. In: Epstein, L., Ferragina, P. (eds.) ESA 2012. LNCS, vol. 7501, pp. 659–670. Springer, Heidelberg (2012)
5. Ma, C., Al-dhelaan, F., Wan, P.-J.: Improved Approximation Algorithm for MISL Under Physical Interference Model, submitted to WASA (2013)
6. Wan, P.-J., Ma, C., Tang, S., Xu, B.: Maximizing Capacity with Power Control under Physical Interference Model in Simplex Mode. In: Cheng, Y., Eun, D.Y., Qin, Z., Song, M., Xing, K. (eds.) WASA 2011. LNCS, vol. 6843, pp. 84–95. Springer, Heidelberg (2011)
7. Wan, P.-J., Xu, X.-H., Frieder, O.: Shortest Link Scheduling with Power Control under Physical Interference Model. In: MSN, pp. 74–78 (2010)

Sweep-Coverage with Energy-Restricted Mobile Wireless Sensor Nodes*

Meng Yang[1], Donghyun Kim[2], Deying Li[1],**,
Wenping Chen[1], Hongwei Du[3,4], and Alade O. Tokuta[2]

[1] School of Information, Renmin University of China, Beijing 100872, China
huashiyangmeng@126.com, {deyingli,chenwenping}@ruc.edu.cn
[2] Department of Mathematics and Computer Science,
North Carolina Central University, Durham NC, USA
{donghyun.kim,atokuta}@nccu.edu
[3] Department of Computer Science and Technology, Harbin Institute of Technology
Shenzhen Graduate School, Shenzhen, China
hongwei.du@ieee.org
[4] Shenzhen Key Laboratory of Internet Information Collaboration, Shenzhen, China

Abstract. Most of the existing results in sweep-coverage focused on minimizing the number of the mobile sensor nodes by carefully planning their corresponding trajectories such that each target of interest can be periodically monitored (within every t time unit). However, the starting locations of the mobile sensors, at which the service depots (or equivalently base stations) of the nodes are usually located, are never considered in the trajectory planning. In order to provide sweep-coverage for a long period of time, each node also needs to periodically visit a base station to replace a battery or refueled (within every T time unit). Motivated by this observation, this paper introduces two new sweep-coverage problems, in which each mobile sensor node is required to visit a base station periodically, namely (t, T)-SCOPe-1 and (t, T)-SCOPe-M, each of which considers one single base station and M base stations for all of the nodes, respectively. We prove those problems are NP-hard and propose heuristic algorithms for them. In addition, we conduct simulations to evaluate the average performance of the proposed algorithms and study their average behavior characteristics.

* This research was jointly supported by National Natural Science Foundation of China under grants 61070191 and 91124001, and the National Research Foundation for the Doctoral Program of Higher Education of China under grant 20100004110001. This work was also supported in part by US National Science Foundation CREST No. HRD-0833184 and by US Army Research Office No. W911NF-0810510. This work received the support from the National Natural Science Foundation of China under Grants 61100191,and by Shenzhen Strategic Emerging Industries Program under Grant No. ZDSY20120613125016389 and JSE201007140001A, and Natural Scientific Research Innovation Foundation in Harbin Institute of Technology under project HIT.NSFIR.2011128.

** Corresponding Author.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 486–497, 2013.

1 Introduction

Nowadays, wireless sensor networks are widely adopted to various surveillance and monitoring systems. In the literature, the *coverage* of a wireless sensor network refers its capability to provide a surveillance service with a certain quality. Given an area (or a set of targets) of interest, a wireless sensor network with *full-coverage* can monitor the whole area (or all of the targets) concurrently [1–3]. Over years, people found full-coverage is an essential requirement of wireless sensor networks in various applications. However, there exist some cases where this coverage model is not ideal. For instance, to protect a border from the intrusion of an enemy, it would be sufficient to have a seamless line of sensor nodes across the border such that any intruder trying to penetrate the sensor field has to be detected. Frequently, such a coverage is referred as *barrier-coverage* [4–10].

Both coverage models discussed above seem different, but they have an important characteristic in common. They require an area or a set of targets of interest to be constantly under surveillance, and assume there are enough sensor nodes to achieve this goal. In [11], Cheng et al. pointed out these existing coverage problems are not applicable to a situation where we have a few number of mobile sensor nodes and a number of targets to be periodically monitored. To address this challenging issue, they introduced a new coverage requirement called *t-sweep-coverage*, whose goal is to make each target periodically observed by a mobile sensor within every t time unit. This new coverage model is especially useful to develop various applications of wireless networks of mobile sensor nodes such as scientific environmental data collection in which the data collected is not time-sensitive. In essence, the sweep-coverage model is unique since the other existing coverage models do not consider how to control mobile sensor nodes to cover the targets.

To the best of our knowledge, the existing studies in sweep-coverage mainly focused on minimizing the number of the mobile sensor nodes by carefully planning their corresponding trajectories such that each target of interest is periodically monitored [11–14]. In many real world scenarios, however, the mobile sensor nodes are expected to be operational with a limited power source. Therefore, in order to provide sweep-coverage for a long period of time, each node also needs to periodically visit a base station to replace a battery or refueled. Due to the reason, it is more desirable to determine the trajectory of each mobile sensor node very carefully when we solve a sweep-coverage problem such that (a) each of the targets of interest is t-sweep-covered for some positive constant t, (b) the number k of mobile sensor nodes is minimized, and (c) each mobile sensor node visits a base station within every T time units for a given positive constant T.

Motivated by this observation, this paper introduces two new sweep-coverage problems, in which each mobile sensor node is also required to visit a base station periodically, namely (t, T)-SCOPe-1 and (t, T)-SCOPe-M, where 1 and M implies the number of base stations available for the nodes in each problem, respectively. Specifically, the formal definition of (t, T)-SCOPe-1 is as below.

Definition 1 ((t, T)-SCOPe-1). *Given a set of targets of interest and a single base station, the* (t, T)-sweep-coverage optimization problem with a base station ((t, T)-SCOPe-1) *is to find the minimum number k of the mobile sensor nodes, which are originally located at the base station, and determine their corresponding trajectories such that each of the targets can be t-sweep-covered and each sensor node visits the base station within every T time units.*

On the other hand, (t, T)-SCOPe-M can be considered as a generalized version of (t, T)-SCOPe-1 with a given constant $M \geq 2$, and its formal definition is as below.

Definition 2 ((t, T)-SCOPe-M). *Given a set of targets of interest and a set of M base stations, the* (t, T)-sweep-coverage optimization problem with M base stations ((t, T)-SCOPe-M) *is to find the minimum number k of the mobile sensor nodes, each of which is initially located at one of the base stations, and determine their corresponding trajectories such that each of the targets can be t-sweep-covered and each sensor node visits its base station within every T time units.*

The contributions of this paper can be summarized as follows: (a) We introduce a new problem, (t, T)-SCOPe-1 and its generalized version, (t, T)-SCOPe-M with sufficient motivation and background for them. (b) We prove the NP-hardness of (t, T)-SCOPe-1 and correspondingly (t, T)-SCOPe-M. We also introduce the *minimum tours with a base station (MinTs1BS)* problem, which is a relaxed version of (t, T)-SCOPe-1, and prove its NP-hardness. (c) We propose a heuristic algorithm for MinTs1BS, namely MinTs1BS-Expand. Based on this algorithm, we design two new heuristic algorithms, SCOPe-1-Solver and SCOPe-M-Solver, for (t, T)-SCOPe-1 and (t, T)-SCOPe-M, respectively. (d) We study the average performance and behavior characteristic of the proposed algorithms via simulation.

The rest of this paper is organized as follows. Section 2 and Section 3 introduce our new algorithms for (t, T)-SCOPe-1 and (t, T)-SCOPe-M, respectively. We present simulation results and make discussions in Section 4. Finally, we conclude this paper and present the future works in Section 5.

2 A New Heuristic Algorithm for (t, T)-SCOPe-1

In (t, T)-SCOPe-1, we assume that there exist a single base station v_b and a set P of m targets of interest, $\{p_1, p_2, \cdots, p_m\}$, deployed on a rectangle region R (see Fig. 1). Each mobile sensor node is initially located at the base station and moves around such that each target can be monitored within every t time units. Each mobile sensor node has a limited energy source, and therefore, it has to revisit the base station within every T time units. We denote the Euclidean distance between two targets p_i and p_j by $dist(p_i, p_j)$. We assume the speed of all mobile sensor nodes is uniformly v. Then, we can denote the maximum length that a mobile sensor can move in T time units by $L = v \times T$. We ignore the sensing range of each node and assume a target is covered by a mobile sensor

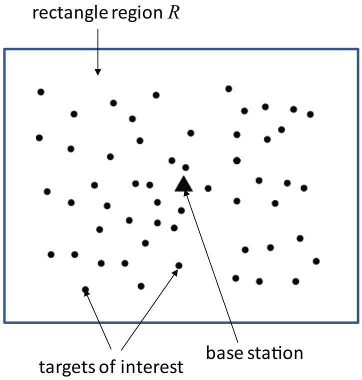


Fig. 1. An illustration of (t, T)-SCOPe-1 problem instance. (t, T)-SCOPe-M assumes $M \geq 2$ base stations.

node only when the mobile sensor node reaches at the target. This assumption is very reasonable if the travel distance of each mobile sensor node is very large. To have a feasible solution, without a loss of generality, we assume $dist(v_b, p_j) \leq \frac{L}{2}$ for any $p_j \in P$. Otherwise, no tour T_i including p_j and v_b such that $C(T_i) \leq L$ can exists, where $C(T_i)$ is the cost of the a graph object T_i, which is defined as the sum of the length of the edges in T_i. We would like to emphasize that while one may think our algorithms are similar to OSWEEP [12], ours are quite different from this algorithm mainly because we do not use a PTAS for TSP, which is computationally intensive, to solve our problems unlike OSWEEP, not to mention the fact that we are dealing with a different problem.

2.1 Computing Minimum Number of Tours Originating from a Base Station

In this section, we study a new problem called the *minimum tours with a base station (MinTs1BS)* problem, and design a heuristic algorithm for it. In the following section, we use this result to solve (t, T)-SCOPe-1.

Definition 3 (MinTs1BS). *Given a positive constant L, a graph $G = (V, E, w)$, and $v_b \in V$ representing the only base station, where w is an edge weight function, the* minimum tours with a base station (MinTs1BS) *problem is to find the minimum number of tours such that (a) each tour includes v_b, (b) each node in $V \setminus \{v_b\}$ is included in some tour, and (c) the cost of each tour does not exceed L.*

Theorem 1. *MinTs1BS is NP-hard.*

Proof. We show MinTs1BS is NP-hard by proving the decision version of MinTs1BS is NP-complete. This can be done by reducing the decision version of TSP, which is a very well-known NP-complete problem, to the decision version

Algorithm 1. MinTs1BS-Expand $(G = (V, E, w), L, v_b)$

```
Set $l \leftarrow 1$ and $V' \leftarrow V \setminus \{v_b\}$.
while $V' \neq \emptyset$ do
    Find $u \in V'$ such that $w(v_b, u)$ is maximum and $w(v_b, u) \leq \frac{L}{2}$, and set $T_l \leftarrow \{v_b, u\}$.
    loop
        Find a node $u'$ in $V'$ with minimum
            $\Delta(T_l, u') = \min_{v_i, v_j \in T_l} \Big\{ w(v_i, u') + w(u', v_j) - w(v_i, v_j) \Big\}$.
        if the cost of $T_l \bigcup \{u'\}$ is greater than $L$ then
            Break /* quit the loop */
        else
            Add $u'$ to $T_l$ as well as remove $u'$ from $V'$.
        end if
    end loop
    $l \leftarrow l + 1$.
end while
Return $l$ and $T_1, T_2, \cdots, T_l$.
```

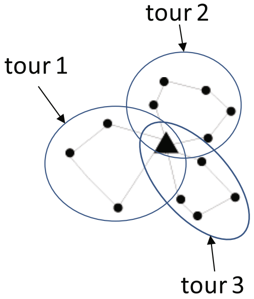


Fig. 2. An example output of MinTs1BS-Expand, which is three node-disjoint tours sharing a single base station

of MinTs1BS as follows. Consider a TSP instance which consists of a set of m sites $U = \{u_1, \cdots, u_m\}$ in a $2D$ plane. Note that the goal of TSP is to find the shortest tour which starts from a site, visits all other sites, and returns back to the originating site. Therefore, the corresponding decision version $\mathcal{X}$ of the TSP instance is, to determine whether there is a tour visiting all the sites and its cost (total edge length) does not exceed a given constant L'. Meanwhile, the decision version $\mathcal{Y}$ of MinTs1BS is to determine whether there is k rooted tours such that all the three conditions in Definition 3 are satisfied for a given k.

Now, given a TSP instance $\mathcal{X} = \langle U, L' \rangle$, we construct an MinTs1BS instance $\mathcal{Y} = \langle G, L, v_b \rangle$ as follow: (a) Randomly pick a node $v \in U$ in $\mathcal{X}$, and assign it as the base station v_b in $\mathcal{Y}$, (b) Copy all nodes $U \setminus \{v\}$ in $\mathcal{X}$ to V in $\mathcal{Y}$. (c) Copy L' in $\mathcal{X}$ to L in $\mathcal{Y}$. (d) Set the weight w for each edge in the graph G in $\mathcal{Y}$ to be their Euclidean distance. (e) Set k to 1 in $\mathcal{Y}$. Then, the answer for $\mathcal{X}$ is yes if and only if that for $\mathcal{Y}$ is yes, which implies that $\mathcal{X}$ and $\mathcal{Y}$ are in fact equivalent. As a result, MinTs1BS is NP-hard.

Next, we propose a heuristic algorithm MinTs1BS-Expand for MinTs1BS. We exploit the strategy used for the algorithm MinExpand in [12], whose goal is to find the minimum number of node-disjoint tours such that the cost of each tour does not exceed L and every node is included in one of the tours. Different from MinExpand, the goal of MinTs1BS-Expand is to find the minimum number of tours such that each tour includes the base station node v_b, the cost of each tour does not exceed L, and all other nodes are included in some tour (see Fig. 2).

MinTs1BS-Expand builds the tours one by one. To construct a new tour T_l, the algorithm first constructs T_l only with v_b. Next, it searches another node $u \in V'$ such that $w(v_b, u) = dist(v_b, u)$ is maximum, but still $w(v_b, u) \leq \frac{L}{2}$ is satisfied, where V' is the set of nodes in V excluding v_b and any node considered so far. From now on, the algorithm iteratively finds a node u' in V' and adds it to T_l such that

$$\Delta(T_l, u') = \min_{v_i, v_j \in T_l} \Big\{ w(v_i, u') + w(u', v_j) - w(v_i, v_j) \Big\}$$

becomes minimum. If $C(T_l) + \Delta(T_l, u') \leq L$, we augment T_l to $T_l \cup \{u'\}$. Otherwise, there is no useful $u' \in V'$, and thus we stop augmenting T_l and continue working on T_{l+1}. This whole process will be repeated until V' becomes empty. Algorithm 1 is the formal description of MinTs1BS-Expand.

Theorem 2 (Lower Bound). *The number of tours generated by* MinTs1BS-Expand *is at least* $\frac{C(T_{mst})}{L}$*, where* T_{mst} *is the minimum spanning tree of all targets of interest and the base station.*

Proof. Consider an output of MinTs1BS-Expand, which is a set of tours $\{T_1, \cdots, T_l\}$ for some l. We first claim

$$C(T_{mst}) \leq \sum_{1 \leq i \leq l} C(T_i).$$

Clearly, this is true since

(a) each T_i is edge disjoint, which implies

$$\sum_{1 \leq i \leq l} C(T_i) = C(\bigcup_{1 \leq i \leq l} T_i), \text{ and}$$

(b) $\bigcup_{1 \leq i \leq l} T_i$ is a graph connecting all targets of interest and the base station, and T_{mst} is a graph with the smallest cost in this kind, which implies

$$C(T_{mst}) \leq C(\bigcup_{1 \leq i \leq l} T_i).$$

Algorithm 2. SCOPe-1-Solver (P, t, T, v_b)

```
$k \leftarrow 0$
Induces a complete graph $G$ of $V = P \bigcup \{v_b\}$.
$\langle l, T_1, \cdots, T_l \rangle \leftarrow$ MinTs1BS-Expand $(G, T \times v, v_b)$.
for each $1 \leq i \leq l$ do
    Follow the tour $T_i$ toward one direction (i.e. counter clockwise) and divide the tour into the minimum number of path segments whose length is no greater than $L_0 = v \cdot t$.
    For each segment generated, we place a mobile sensor node at the beginning point of the segment, and let the mobile node moves following $T_i$ (i.e. counter clockwise).
    $k \leftarrow k + \lceil \frac{C(T_i)}{L_0} \rceil$.
end for
Return $k$ and the sub-tours for the mobile sensor nodes.
```

In addition, by the definition of the problem, in any feasible solution, $C(T_i) \leq L$ has to be true. As a result, we have

$$C(T_{mst}) \leq l \cdot L \Leftrightarrow \frac{C(T_{mst})}{L} \leq l.$$

2.2 A New Heuristic Algorithm for (t, T)-SCOPe-1

In this section, we propose a new heuristic algorithm for (t, T)-SCOPe-1, namely SCOPe-1-Solver, based on MinTs1BS-Expand introduced in the previous section. Before proceeding any further, let us discuss about the NP-hardness of (t, T)-SCOPe-1 first by introducing the following theorem.

Theorem 3. *(t, T)-SCOPe-1 is NP-hard.*

Proof. Consider a special case of (t, T)-SCOPe-1 with $t = L$ and $T = L$. Then, (t, T)-SCOPe-1 is reduced to MinTs1BS, which is proven to be NP-hard in Theorem 1.

Next, let us discuss about SCOPe-1-Solver in detail. Given a (t, T)-SCOPe-1 problem instance $\langle P, t, T, v_b \rangle$, the algorithm first induces a complete graph G of $V = P \cup \{v_b\}$. For any two nodes in V, the weight of the edge between them is the Euclidean distance between them. Then, $\langle G, L \leftarrow T \times v, v_b \rangle$ is an MinTs1BS problem instance. Next, SCOPe-1-Solver applies MinTs1BS-Expand to this instance and obtain l tours, $T_1, T_2, \cdots, T_l$, each of which includes v_b.

Now, for each tour T_i, we follow the tour toward one direction (i.e. counter clockwise) and divide T_i into a minimum number of path segments whose length is no greater than $L_0 = v \cdot t$. As a result, we will have $\lceil \frac{C(T_i)}{L_0} \rceil$ path segments. For each segment, at the beginning of the segment, we place a mobile sensor node. Then, we let the $\lceil \frac{C(T_i)}{L_0} \rceil$ mobile sensor nodes move following T_i (i.e. counter clockwise). In this way, each mobile sensor node will visit the base station within

Algorithm 3. SCOPe-M-Solver (P, t, T, B)

```
$k \leftarrow 0$
for each base station $v_{b_i} \in B$ such that $|B| = M$ do
    Set a cluster $C_i \leftarrow \{v_{b_i}\}$.
end for
for each target $p_j$ in $P$ do
    Add $p_j$ to $C_i$, where $v_{b_i}$ is the nearest base station of the target.
end for
for each cluster $C_i$ do
    $k \leftarrow k$ + SCOPe-1-Solver $(C_i, t, T, v_{b_i})$.
end for
Return $k$ and the sub-tours for the mobile sensor nodes (determined by SCOPe-1-Solver).
```

every T unit time and each target on T_i will be visited by a mobile sensor node at least every t unit time. Algorithm 2 is the formal definition of SCOPe-1-Solver.

3 A New Heuristic Algorithm for (t, T)-SCOPe-M

Next, we consider (t, T)-SCOPe-M. Note that as shown by the theorem below, this problem is also NP-hard.

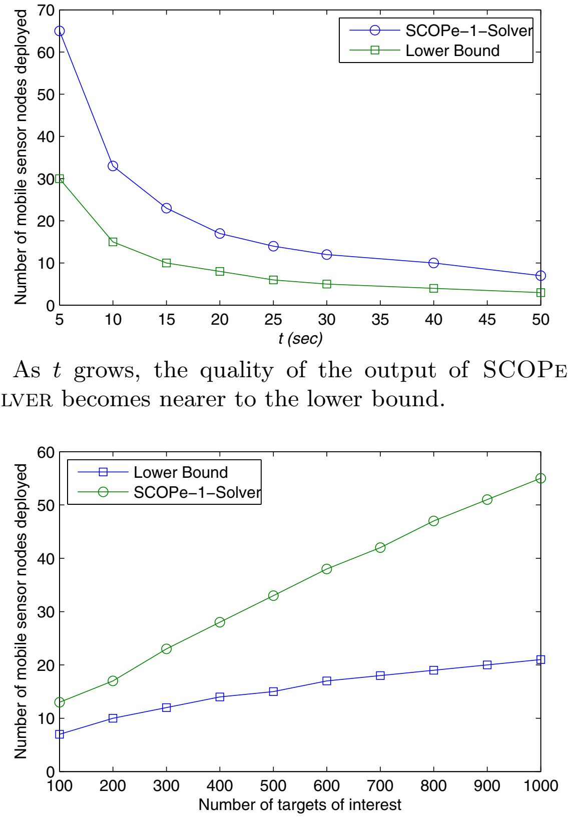
Theorem 4. *(t, T)-SCOPe-M is NP-hard.*

Proof. By Theorem 3, (t, T)-SCOPe-1 is NP-hard. Since (t, T)-SCOPe-1 is a special case of (t, T)-SCOPe-M, (t, T)-SCOPe-M is also NP-hard. Therefore, this theorem is true.

Due to the reason, we propose SCOPe-M-Solver, a heuristic algorithm for (t, T)-SCOPe-M, namely SCOPe-M-Solver. Remind that unlike (t, T)-SCOPe-1, (t, T)-SCOPe-M assumes M base stations, each of which can service one or more mobile sensor nodes. The main idea of SCOPe-M-Solver is partitioning the network into multiple clusters, each of which includes one base station and a subset of targets. Then, each cluster can be treated as a (t, T)-SCOPe-1 problem instance. In detail, given a set of problem input parameters P, t, T, and $B = \{v_{b_1}, \cdots, v_{b_M}\}$ (the set of base stations), we first assign each target in P as a member of the closest base station in B. In this way, we obtain a set of clusters $\{C_1, C_2, \cdots, C_M\}$, each of which is with one base station and its members (targets). Next, for each cluster C_i, we apply SCOPe-1-Solver and assign mobile sensor nodes. Algorithm 3 is the formal definition of SCOPe-M-Solver.

4 Simulation Results and Analysis

Now, we present our simulation results, evaluate the average performance of SCOPe-1-Solver and SCOPe-M-Solver, and study their average behavioral characteristics.



(a) As t grows, the quality of the output of SCOPe-1-Solver becomes nearer to the lower bound.

(b) Our simulation results indicate that the solution generated by SCOPe-1-Solver may have the redundancy of up to 60 percent.

Fig. 3. The simulation results indicate SCOPe-1-Solver has a decent performance, but there is still a good chance to be improved

4.1 Performance Analysis of SCOPe-1-Solver

In this section, we compare the performance of SCOPe-1-Solver against the lower bound introduced in Theorem 2 (the lower bound of of MinTs1BS is also the lower bound of (t, T)-SCOPe-1), which is $\frac{C(T_{mst})}{L} = \frac{C(T_{mst})}{v \cdot t}$, where $C(T_{mst})$ is the cost of the minimum spanning tree T_{mst} over the all targets and the base station. The performance comparison is conducted throughout two difference settings.

In the first setting, we prepare a 100m by 100m virtual space R and set $v = 10$m/sec, $T = 100$sec. Then, we vary t from 5sec to 50sec increased by 5sec. For each parameter setting, we randomly deploy 500 targets. Fig. 3(a) presents the performance of SCOPe-1-Solver against the lower bound. From the figure, we can learn that

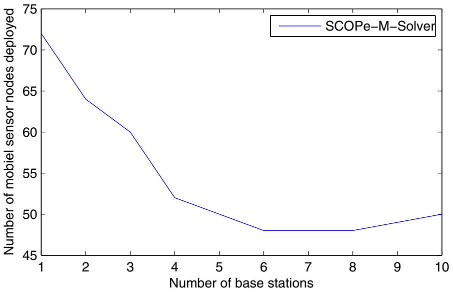


Fig. 4. This result shows that the number of base stations has a limited impact over the number of mobile sensor nodes employed by SCOPe-M-Solver.

(a) both algorithms deploy less number of sensor nodes as t increases, which is easy to expect, and

(b) roughly, our algorithm deploys twice more nodes than the lower bound.

The second setting is similar to the first one, but we fix $t = 10$sec and vary the number of targets from 100 to 1000 increased by 100. Fig. 3(b) shows the average number of mobile sensor nodes deployed by SCOPe-1-Solver against the lower bound. Our simulation results indicate that the solution generated by SCOPe-1-Solver may have a certain amount of redundancy than that of the lower bound (since the lower bound can be not tight and the optimal solution can be significantly greater than the lower bound). This means that we may have a chance to reduce the number of mobile sensor nodes deployed by SCOPe-1-Solver by reorganizing the subset of nodes covered by each mobile sensor node.

From the simulation results, we can observe that as the number of targets increases, the difference between the cost of our algorithm and the lower bound grows. We claim that this happens due to our choice of the lower bound which is not necessarily tight. In detail, consider an optimal solution of (t, T)-SCOPe-1 whose cost is always greater than or equal to the lower bound. Given T and t, it is clearly true that an optimal solution of the problem will induce more number of tours rooted at the base station as the number of targets increases. Note that each tour is basically a set of edges spanning over a subset of targets plus two edges each of which is from one of the targets to the base station. Also, as the number of targets increases, more number of such edges connecting targets to the base station exist in the optimal solution. Meanwhile, we simply divide the length of an MST connecting all targets and the base station by L to obtain the lower bound. As a result, the computation of lower bound only considers two edges from a target to a base station.

In conclusion, our simulation results indicate SCOPe-1-Solver has a decent performance especially with larger t, but there is still a good chance to improve the algorithm and the lower bound.

4.2 Performance Analysis of SCOPe-M-Solver

Now, we study the impact of the number M of base stations over the number of mobile sensors deployed by SCOPe-M-Solver. In this simulation, we prepare a 100m × 100m virtual space and randomly deploy 1600 targets of interest. We also set $v = 10$m/sec, $T = 100$sec, and $t = 10$sec.

Fig. 4 illustrates our simulations result. Very interestingly, the simulation results indicate that given a set of targets of interest, increasing the number of base stations, which tend to be very expensive, is not always beneficial, but there is a cost-effective number of base stations. As a result, SCOPe-M-Solver could be helpful to reduce the construction and operation cost of a wireless sensor network supporting sweep-coverage only in some extent. Still, SCOPe-M-Solver seems to produce a cost-effective solution if the wireless sensor network operates over a long time period and the operation cost of each sensor node is the major operation cost of each base station.

5 Conclusions

In this paper, we introduced two new sweep-coverage problems in wireless sensor networks. We showed the problems are NP-hard and proposed heuristic algorithms for them. We also conducted simulations to study the average performance and behavioral characteristics of the proposed algorithms. Our simulations result indicate that the proposed algorithms have a decent performance on average but still there is a good room to improve their performance. Therefore, as a future work, we plan to further investigate better solutions for the proposed problems. This paper made several assumptions such as uniform speed of each mobile node. In the near future, we also plan to study more realistic version of the problems with less number of assumptions.

References

1. Wan, P., Yi, C.: Coverage by Randomly Deployed Wireless Sensor Networks. IEEE Transactions on Information Theory 52(6), 2658–2669 (2006)
2. Bai, X.: Deploying Four-connectivity and Full-coverage Wireless Sensor Networks. In: Proc. of the 27th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2008), Phoenix, AZ (2008)
3. Huang, C.F., Tseng, Y.C.: The Coverage Problem in a Wireless Sensor Network. In: Proc. of the 2nd International Workshop on Wireless Sensor Networks and Applications (WSNA 2003), San Diego, CA (2003)
4. Kumar, S., Lai, T.H., Arora, A.: Barrier Coverage with Wireless Sensors. In: Proc. of the 11th Annual International Conference on Mobile Computing and Networking (MobiCom 2005), Cologne, Germany (2005)
5. Kim, D., Kim, J., Li, D., Kwon, S.-S., Tokuta, A.O.: On Sleep-wakeup Scheduling of Non-penetrable Barrier-coverage of Wireless Sensors. In: Proceedings of the IEEE Global Communications Conference (GLOBECOM 2012) (December 2012)

6. Liu, B., Dousse, O., Wang, J., Saipulla, A.: Strong Barrier Coverage of Wireless Sensor Networks. In: Proc. of the 9th ACM Interational Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2008), Hong Kong, China (2008)
7. Yang, G., Qiao, D.: Barrier Information Coverage with Wireless Sensors. In: Proc. of the 28th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2009), Rio de Janeiro, Brazil (2009)
8. Saipulla, A., Westphal, C., Liu, B., Wang, J.: Barrier Coverage of Line-based Deployed Wireless Sensor Networks. In: Proc. of the 28th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2009), Rio de Janeiro, Brazil (2009)
9. Yang, G., Qiao, D.: Multi-round Sensor Deployment for Guaranteed Barrier Coverage. In: Proc. of the 29th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2010), San Diego, CA (2010)
10. Chen, A., Li, Z., Lai, T., Liu, C.: One-way Barrier Coverage with Wireless Sensors. In: Proc. of the 30th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2011), Shanghai, China (2011)
11. Cheng, W., Li, M., Liu, K., Liu, Y., Li, X.Y., Liao, X.: Sweep Coverage with Mobile Sensors. In: Proc. of IEEE International Parallel & Distributed Processing Symposium (IPDPS 2008), Miami, FL (2008)
12. Du, J., Li, Y., Liu, H., Sha, K.: On Sweep Coverage with Minimum Mobile Sensors. In: Proc. of the 2010 IEEE 16th International Conference on Parallel and Distributed Systems (ICPADS 2010), Shanghai, China, pp. 283–290 (2010)
13. Zhao, D., Ma, H., Liu, L.: Mobile Sensor Scheduling for Timely Sweep Coverage. In: 2012 IEEE Wireless Communications and Networking Conference (WCNC 2012), Paris, France (2012)
14. Cheng, T.M., Savkin, A.V., Javed, F.: Decentralized Control of a Group of Mobile Robots for Deployment in Sweep Coverage. Robotics and Autonomous Systems 59(7-8), 497–507 (2011)

The Trading between Virtual Mobile Operator and Wireless Service Provider in the Two-Tier Femtocell Network

Fan Yang[1], Bo Yang[1,2,⋆], and Xinping Guan[1]

[1] Department of Automation, Shanghai Jiao Tong University
and Key Laboratory of System Control and Information Processing,
Ministry of Education of China
[2] Key Laboratory of Education Ministry for Image Processing and Intelligent Control
{zerozyt,bo.yang,xpguan}@sjtu.edu.cn

Abstract. This paper considers the problem that the virtual mobile operator (VMO) buys services from the wireless service provider (WSP) to serve its users. When faced with poor indoor coverage or at the edge of marcocell, the WSP would ask nearby femto base stations (FBSs) to help serve the VMO. We focus on the interesting questions when the WSP prefers to ask help from FBSs and how WSP, VMO and FBSs adjust their strategies to maximize their profits. A two-stage Stackelberg game is built for the situation when the WSP serve the VMO by itself and a four-stage Stackelberg game is considered when the WSP rents FBSs by offering spectrum bands to FBSs. We firstly define the utility functions of VMO, WSP and FBSs to reflect their satisfaction and cost in participating into the game and then give the analysis based on the backward induction method. Meanwhile, the competition among FBSs when the WSP allocates spectrum bands for FBSs' help is formulated as a non-cooperative spectrum competition game (NSCG) and the existence and uniqueness of the Nash equilibrium in NSCG is proved. Simulation results of the two service modes show the WSP can obtain more profits through renting FBSs in poor coverage areas.

Keywords: Femtocell Networks, Stackelberg Game, Pricing, Spectrum allocation.

1 Introduction

In recent years, a demand for better quality of service (QoS) and higher data rates is steadily increasing because of the constant growth of mobile terminals. By 2017, global mobile data traffic will reach 11.2 exabytes per month and increase 13-fold from 2012 to 2017 [1]. Enjoying software services provided by companies like Facebook or YouTube through mobile terminals becomes more and more popular. However, it is hard for these companies, called virtual mobile operator (VMO), to serve their users by themselves because they own no

⋆ Corresponding author.

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 498–511, 2013.

spectrum resource or macro base station (MBS). On the contrary, the wireless service provider (WSP), like Verizon Wireless or AT&T Wireless, has the license of spectrum and MBS. In order to serve VMO's users (VUs), the VMO would like to buy services from the WSP, and the WSP is willing to serve the VMO for extra income.

Besides, users receive poor signal when they are indoors or at the edge of macrocell [2]. Femtocell base stations (FBSs) have been proposed as an effective way to solve this problem. Because of the proximity between transmitting ends and receiving ends, the indoor femtocell users or those at the edge of marcocell experience better quality of service and higher data rates when they can be connected to the nearby FBSs [3]. Unlike the WSP, FBSs do not own the licensed spectrum so if the WSP wants them to help itself, it should give them part of its spectrum bands as exchange. Hence the WSP should take both the spectrum resources it pays and better services it earns into consideration. Weighing the benefits, the WSP will determine to ask FBSs for help or not.

Stackelberg games are widely used to tackle multiple-tier network problem. In [4], Stackelberg game is used to model and analyze the interactions between the cognitive mobile virtual network operator and secondary unlicensed users, while in [5] the authors study the energy efficiency aspect of spectrum sharing and power allocation in heterogeneous cognitive radio networks with femtocells. In [6] a three-stage Stackelberg game is built to describe the scenario where a femtocell service provider expects to rent spectrum from the coexisting macrocell service provider to serve its end users. In this paper, we build a two-stage Stackelberg game to describe the situation where the WSP serves the VMO by itself. And then we move forward to set up a four-stage Stackelberg game to describe the condition that WSP rents FBSs to help serve VMO. The backward induction is used to solve these two Stackelberg games.

There are three types of access policies for femtocell: open-access (OA) opening for all users, closed-access (CA) opening for subscriber users only and hybrid-access (HA) opening partly for non-subscriber users [7]. Previous works focus on CA considering power control in [8] [9]. In [10], the WSP determines the reward to femtocells offering to help offload proportionally to the open ratios of FBSs. [11] describes a case where FBSs compete with each other to decide whether to use a closed or an open access policy and finally results in a unique Nash equilibrium. When the WSP selects to ask FBSs for help, firstly we focus on maximizing every FBS's profits and suppose that FBSs use the HA method by constantly adjusting the open ratio at the fourth stage of the four-stage Stackelberg game. Secondly at the third stage the WSP will determine the reward proportionally to the throughput that every FBS provides.

The outline of this paper is as follows. In Section 2 and Section 3, we build the system model and provide the problem formulation respectively. In Section 4, the formulated Stackelberg games are solved using backward induction method. In Section 5, simulation results and numerical analysis are given. Conclusions are presented in Section 6.

2 System Model

In this section, we will introduce the system model in the context of the VMO served by the WSP with and without FBSs. Fig.1 shows the scenario that the VMO trades with the WSP and FBSs.

2.1 Service without FBSs

Companies called the VMO often need to provide services to a huge number of VUs. However, the VMO does not own licensed spectrum or MBSs. The WSP is willing to provide the VMO spectrum bands and marcocell for extra income. In this paper, the typical model that one VMO trades with one WSP is considered, and it will be extended to multi-trader scenario in the future.

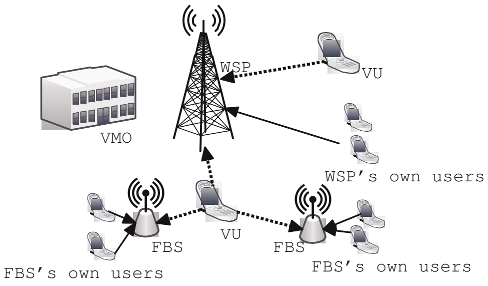


Fig. 1. The WSP serves the VUs or rents FBSs to help serve

The set of all VUs is denoted by $\mathcal{J} = \{1, ..., j, ..., J\}$. In this scenario, we suppose that the WSP serves the whole set $\mathcal{J}$ by itself in a TDMA manner. We suppose the WSP has total W units of spectrum and serves VMO with m_W units of spectrum. For convenience, we suppose every VU has the same downlink Signal to Noise Ratio (SNR) η_W when it connects to the WSP's MBS, which can be implemented by adjusting the transmission power of MBS. Considering the transmission time λ_j in a unit frame for VU j such that $\lambda_{W1} + ... + \lambda_{Wj} + ... + \lambda_{WJ} = 1$, the data rate of VU j served by the WSP is:

$$R_{Wj} = \lambda_{Wj} \log_2(1 + \eta_{Wj}), \tag{1}$$

where $\eta_{Wj} = \frac{h_{Wj}^2 P_{Wj}}{\sigma^2}$ represents the SNR between the WSP and VU j. h_{Wj}^2 is the channel gain between the WSP and VU j. The transmission power of the WSP to VU j is P_{Wj}, and the variance of complex Gaussian thermal noise is denoted by σ^2. Since $\eta_{W1} = ... = \eta_{Wj} = ... = \eta_{WJ} = \eta_W$ (through adjusting the transmission power P_{Wj}), we have the sum data rate of VUs served by WSP $R_W = \log_2(1 + \eta_W)$.

The better services are, the more satisfied users will be. The satisfaction of the VMO or the WSP is related to the throughput that its users receive. We propose a concave function of total throughput to represent the satisfaction. Such function should be monotonically increasing because the more throughput VMO or WSP has, the more satisfied it will be. And the satisfaction value should approach almost a fixed value when the throughput approach infinite. As a result, the satisfaction functions of the WSP and the VMO are expressed respectively as follows:

$$S_W = A_W \left(1 - e^{-a_W (W - m_W) R_W}\right), \tag{2}$$

$$S_V = A_V \left(1 - e^{-a_V m_W R_W}\right), \tag{3}$$

where A_W and A_V are the adjustment coefficients. a_W and a_V denote the coefficients of satisfaction sensitivity such that higher value means less sensitivity. m_W is the amount of spectrum that the WSP uses to serve VUs. $m_W R_W$ is the throughput the WSP provides to the VMO and $(W - m_W)R_W$ is the throughput the WSP leaves to its own users.

We assume the total money the VMO pays to the WSP is:

$$M = m_W R_W p, \tag{4}$$

where p denotes the price per unit throughput.

2.2 Service with FBSs

When faced with poor indoor coverage or at the edge of marcocells, the WSP will ask some FBSs for help to enhance its coverage. In this paper, a few FBSs are assumed to be deployed sparsely (there is no inter-interference between FBSs) and randomly by Poisson distribution. We suppose, in poor coverage areas, the WSP rents FBSs by paying FBSs with spectrum band w. The set of FBSs is $\mathcal{F} = \{F_1, ..., F_i, ..., F_I\}$ and $\dot{I}$ of these FBSs ($\dot{\mathcal{F}} = \{F_1, ..., F_i, ..., F_{\dot{I}}\}$) are supposed to be willing to help the WSP serve VUs. We suppose the set of VUs served by the WSP is $\overline{\mathcal{J}_W} = \{1, ..., j, ..., \overline{J_W}\}$ (the bar "-" on the head of a variable means it is in the situation that the WSP ask FBSs for help), and the set of VUs served by F_i is $\mathcal{J}_{F_i} = \{1, ..., j, ..., J_{F_i}\}$. For simplicity, we assume that one VU can only be served by one FBS or the WSP and all of the VUs have the same downlink SNR η_F when they connect to the same FBS F_i or the downlink SNR η_W (the same as mentioned in the last subsection) when they connect to the WSP by some proper power control. The interference among VUs served by the same FBS or the WSP is eliminated by obeying the TDMA scheduler. Considering the transmission time $\lambda_{F_i j}$ of VU j ($\lambda_{F_i 1} + ... + \lambda_{F_i j} + ... + \lambda_{F_i J_{F_i}} = 1$), the data rate between FBS F_i and VU j can be expressed as:

$$R_{F_i j} = \lambda_{F_i j} \log_2(1 + \eta_{F_i j}), \tag{5}$$

where $\eta_{F_ij} = \frac{h_{F_ij}^2 P_{F_ij}}{\sigma^2}$, $h_{F_ij}^2$ is the channel gain between the FBS F_i and VU j, and P_{F_ij} denotes the transmission power of FBS F_i to VU j. Since $\eta_{F_i1} = ... = \eta_{F_ij} = ... = \eta_{F_iJ} = \eta_{F_i}$, we have the sum data rate of VUs served by FBS F_i $R_{F_i} = \log_2(1 + \eta_{F_i})$.

Considering the transmission time $\overline{\lambda_{Wj}}$ of VU j ($\overline{\lambda_{W1}}+...+\overline{\lambda_{Wj}}+...+\overline{\lambda_{W\overline{J_W}}} = 1$), then the data rate between the WSP and VU j can be expressed as:

$$\overline{R_{Wj}} = \overline{\lambda_{Wj}} \log_2(1 + \overline{\eta_{Wj}}), \tag{6}$$

where $\overline{\eta_{Wj}} = \frac{\overline{h_{Wj}}^2 \overline{P_{Wj}}}{\sigma^2}$ represents the SNR between the WSP and VU j. $\overline{h_{Wj}}^2$ is the channel gain between the WSP and VU j. The transmission power of the WSP to VU j is $\overline{P_{Wj}}$. Since $\overline{\eta_{W1}} = ... = \overline{\eta_{Wj}} = ... = \overline{\eta_{WJ}} = \eta_W$, we have the sum data rate of VUs served by WSP $\overline{R_W} = R_W = \log_2(1 + \eta_W)$.

It is supposed that there is no inter-interference between the WSP and FBSs since they can use different spectrum bands.

Like (2) and (3), we give the FBS's satisfaction function:

$$S_{F_i} = A_F \left(1 - e^{-a_F(w_i - m_{F_i})R_{F_i}}\right), \tag{7}$$

where all of the FBSs have the same A_F and a_F. m_{F_i} is the spectrum that FBS F_i uses to help serve VUs and w_i is the spectrum offered by the WSP to FBS F_i. Note that $w = \sum_{i=1}^{I} w_i$ and $(wi - m_{F_i})$ denotes the spectrum that FBS F_i leaves to its own subscriber.

The satisfaction functions of WSP and VMO in the presence of FBSs are given respectively:

$$\overline{S_W} = A_W \left(1 - e^{-a_W(W - \overline{m_W} - w)R_W}\right), \tag{8}$$

$$\overline{S_V} = A_V \left(1 - e^{-a_V(\overline{m_W}R_W + \sum_{i=1}^{I}(m_{F_i}R_{F_i}))}\right), \tag{9}$$

where $(W - \overline{m_W} - w)R_W$ is the throughput left to the WSP. $\overline{m_W}R_W + \sum_{i=1}^{I}$ $(m_{F_i}R_{F_i})$ denotes the sum throughput provided by WSP and FBSs.

In the considered scenario, the FBSs are deployed by different end users and their operator is not the WSP in Fig.1, so FBSs will compete with each other. The WSP should refund FBSs according to the throughput they provide. We suppose that the WSP divides the spectrum band w proportionally to the throughput that every FBS provides:

$$w_i = \frac{m_{F_i}R_{F_i}}{\sum_{i=1}^{I} m_{F_i}R_{F_i}} w. \tag{10}$$

Notice that the VMO does not realize the existence of service from FBSs. It supposes the total throughput is provided by WSP. So the VMO pays the WSP for the service with unit price $\overline{p}$:

$$\overline{M} = (\overline{m_W}R_W + \sum_{i=1}^{I}(m_{F_i}R_{F_i}))\overline{p}. \tag{11}$$

It is reasonable that the VMO pays the WSP according to the sum throughput it receives. However, the spectrum that the WSP loses is more than spectrum that the VMO receives since parts of the former are used by FBSs to serve their own subscribers. Since throughput is equal to spectrum multiplied by data rate, the WSP chooses to ask FBSs for help only when R_W is low enough.

3 Problem Formulation

In this section, we consider two situations. In the first situation, the WSP serves the VMO by itself. While in the second situation, the WSP asks FBSs to help serve the VMO because of its poor coverage in some areas.

3.1 Service without FBSs

In this case, the WSP serves the VMO by itself. We define utility functions respectively to represent profits of the WSP and the VMO.

a) The Utility Function of the VMO. The VMO wants as much spectrum as possible to increase VUs' throughput. However, buying more spectrum means paying more money to the WSP, which will lower VMO's profits. The utility function of VMO is given below to reflect its profit:

$$U_V = S_V - M, \tag{12}$$

where $S_V = A_V \left(1 - e^{-a_V m_W R_W}\right)$ is given in (3), M is given in (4), and A_V is the weighting factor of VMO's satisfaction in U_V.

b) The Utility Function of the WSP. The WSP tends to lease as much spectrum as possible to the VMO in order to earn more money. However, if WSP leases too much spectrum to help the VMO, its satisfaction will decrease. The utility function of WSP is given below:

$$U_W = S_W + M, \tag{13}$$

where S_W is given in (2) and M is given in (4).

3.2 Service with FBSs

In this situation, FBSs help the WSP serve VUs.

a) The Utility Function of the VMO. Like the situation that the WSP serves VUs without FBSs, the VMO cares about both total VUs' throughput and the money it pays. The utility function is written as following:

$$\overline{U_V} = \overline{S_V} - \overline{M}, \tag{14}$$

where $\overline{S_V}$ is given in (9) and $\overline{M}$ is given in (11).

b) The Utility Function of the WSP. As to the WSP's profit, the WSP should take both its remaining throughput and the money it earns into consideration. The utility function of WSP is written as below:

$$\overline{U_W} = \overline{S_W} + \overline{M}, \tag{15}$$

where $\overline{S_W}$ is given in (8) and $\overline{M}$ is given in (11).

c) The Utility Function of the FBSs. Because there is no monetary transaction between the WSP and FBSs, it is reasonable that the utility function only concerns FBS's satisfaction. The utility function of FBS F_i is simply written as

$$U_{Fi} = S_{F_i}, \tag{16}$$

where S_{F_i} is given in (7).

4 Stackelberg Game Analysis

In this section, we set up two Stackelberg games and analyze them by backward induction method.

4.1 Service without FBSs

In this two-stage Stackelberg game, the WSP, the leader, announces the price p per unit throughput at the beginning. Following the WSP's action, the VMO decides how much throughput it buys from WSP to maximize its utility value.

a) Best strategy of VMO on spectrum that it leases. Using the backward induction, we firstly analyze the amount of spectrum bandwidth that VMO requires. The best strategy of VMO is determined after the price announced by the WSP.

The target of VMO is to maximize U_V with the constraint $0 \leq m_W \leq W$.

Theorem 1. *The best response strategy of VMO is calculated as:*

$$m_W^*(p) = \frac{\ln(A_V a_V) - \ln p}{R_W a_V}. \tag{17}$$

Proof. See technical report [12].

Given the price p, the VMO can determine m_W^* to maximize its utility.

b) Best strategy of WSP on price setting. From Theorem 1, the VMO can get the best strategy of the leased spectrum. According to the $m_W^*(p)$, the WSP can determine its best strategy in the pricing subgame. The WSP's goal is to maximize the utility value defined in (13). By transforming (17), we obtain:

$$p = e^{\ln A_V a_V - R_W a_V m_W} \tag{18}$$

with the constraint $e^{\ln A_V a_V - R_W a_V W} \leq p \leq A_V a_V$ due to $0 \leq m_W \leq W$.

Theorem 2. *The best strategy of WSP on pricing the throughput is given as:*

$$p^* = \max\{e^{\ln A_V a_V - R_W a_V W}, min\{p_0, A_V a_V\}\}, \tag{19}$$

where $p_0 = \arg(\frac{a_W A_W}{a_V p} e^{-a_W R_W (W - \frac{\ln(A_V a_V)}{R_W a_V} + \frac{\ln p}{R_W a_V})} + \frac{\ln(A_V a_V)}{a_V} - \frac{\ln p}{a_V} - \frac{1}{a_V} = 0)$.

Proof. See technical report [12].

However, there is no analytical solution for p^*. We will give the numerical solution in the simulation part using numerical method like Newton method.

4.2 Service with FBSs

We build a four-stage Stackelberg game to solve this problem. In the first stage, the WSP determines the price p per unit throughput in order to maximize its utility. As the follower, the VMO determines throughput $(\overline{m_W} R_W + \sum_{i=1}^{\dot{I}} m_{F_i} R_{F_i})$ it wants in the second stage. In the third stage, the WSP rents FBSs. In order to maximize its utility, the WSP needs to determine the number of spectrum bands w offered to FBSs. Following the action of WSP, every FBS determines throughput $m_{F_i} R_{F_i}$ it provides to the WSP.

a) The best strategy of FBSs on spectrum competition. Since a specific FBS's profits do not only depend on its option but also subject to other FBSs' actions, it should adjust its strategy considering other FBS's strategies. We form a non-cooperative spectrum competition subgame (NSCG) expressed as $G = (\{F_i\}, \{m_{F_i}\}, \{U_{Fi}(\cdot)\})$ among FBSs. $\{m_{F_i}\}$ is the pure strategy set of F_i, and $\{U_{Fi}(\cdot)\}$ denotes the set of FBSs' utility functions.

Theorem 3. *There exists a unique Nash Equilibrium in the NSCG and the best strategy of F_i is given below:*

$$m^*_{F_i} = \begin{cases} 0, & R_{F_i} \le \frac{C}{w} \\ \frac{T - \frac{(T)^2}{R_{F_i} w}}{R_{F_i}}, & otherwise \end{cases} \tag{20}$$

where $C = \sum_{n=1, n \neq i}^{I} R_{Fn} m_{Fn}$ *and* $T = \sum_{i=1}^{I} R_{F_i} m_{F_i}$.

Proof. See technical report [12].

If $m_{F_i} = 0$, F_i is unwilling to serve the VMO because there is no benefit. Then F_i quits NSCG and the competition among the remaining FBSs forms a new NSCG. It will repeat until no FBS wants to quit. We suppose $\dot{I}$ remaining FBSs serve the VMO and the set of FBSs is updated to $\dot{\mathcal{F}} = \{F_1, ..., F_{\dot{I}}\}$.

b) The best strategy of WSP on spectrum offering to FBSs. At the third stage, $\sum_{i=1}^{\dot{I}} m_{F_i} R_{F_i}$ is fixed. To maximize its utility, the WSP adjusts w given the price $\overline{p}$ and the VMO's demand of throughput m ($m = \overline{m_W} R_W + \sum_{i=1}^{\dot{I}} m_{F_i} R_{F_i}$).

Theorem 4. *The best strategy of WSP is*

$$w^* = \frac{\ln \frac{X_M \overline{p}}{A_W a_W R_W}}{a_W R_W} + W - \overline{m_W} \tag{21}$$

where $X_M = \frac{N-1}{\sum_{i=1}^{\dot{I}} \frac{1}{R_{F_i}}}$.

Proof. See technical report [12].

Given the price per unit throughput and spectrum bands that the VMO buys, the WSP decides w according to (21).

c) The best strategy of VMO on throughput. In the second stage, the VMO will decide how much throughput m it wants from the WSP and FBSs. Because the throughput that FBSs provide has been obtained, the VMO only needs to determine spectrum bands $\overline{m_W}$ ($0 \le \overline{m_W} \le W - w$) leased from the WSP.

In order to simplify notation, let $X_W = \frac{\ln \frac{X_M \overline{p}}{A_W a_W R_W}}{a_W R_W} + W$.

The VMO wants to maximize its utility with the constraint $0 \le \overline{m_W} \le W - w$.

Theorem 5. *The best strategy of VMO is*

$$\overline{m_W}^* = \frac{\ln \frac{\overline{p}}{A_V a_V} + X_M X_W a_V}{(X_M - R_W) a_V}. \tag{22}$$

Proof. See technical report [12].

Given the price per unit throughput, the VMO can decide $\overline{m_W}^*$.

d) The best strategy of WSP on pricing. In this part, the WSP should determine the price $\overline{p}$ ($\overline{p} \ge 0$) for maximizing $\overline{U_W}$.

For convenience, we define $Y = \ln \frac{X_M}{A_W a_W R_W}$.

Combining the constraints $w \ge 0$, $\overline{m_W} \ge 0$ and $w + \overline{m_W} \le W$ with (21) and (22), we can obtain the constraints on $\overline{p}$:

$$\begin{cases} p_{m_W} \le \overline{p} \le \min\{p_w, \dfrac{A_W a_W \overline{R_W}}{X_M}\}, \quad X_M - \overline{R_W} < 0, & \text{(23a)} \\ p_w \le \overline{p} \le \min\{p_{m_W}, \dfrac{A_W a_W \overline{R_W}}{X_M}\}, \quad \textit{otherwise} & \text{(23b)} \end{cases}$$

where $p_w = e^{\frac{a_W R_W \ln A_V a_V + X_M a_V \ln A_W a_W R_W - X_M a_V \ln X_M - a_W R_W X_M a_V W}{a_W R_W + X_M a_V}}$,

and $p_{m_W} = e^{\frac{a_V^2 a_W R_W \ln A_V a_V - a_W a_V R_W^2 W - a_V R_W \ln X_M + a_V R_W \ln A_W a_W R_W}{R_W a_V + a_W R_W}}$.

When calculating p_w and p_{m_W}, we get
$\frac{(a_W R_W + X_M a_V)\ln\overline{p} + X_M a_V \ln X_M - a_W R_W \ln A_V a_V - X_M a_V \ln A_W a_W R_W + a_W R_W X_M a_V W}{(X_M - R_W) a_V a_W R_W}$
≥ 0
and $\frac{\ln p}{R_W - X_M} \geq \frac{a_V^2 a_W R_W \ln A_V a_V - a_W a_V R_W^2 W - a_V R_W \ln X_M + a_V R_W \ln A_W a_W R_W}{(R_W - X_M)(R_W a_V + a_W R_W)}$.
If $X_M - \overline{R_W} < 0$, we get (23a). Otherwise, we get (23b).

Theorem 6. *The best strategy of WSP is*

$$\overline{p}^* = \begin{cases} \max\{p_{m_W}, \min\{\overline{p_0}, p_w, \frac{A_W a_W \overline{R_W}}{\overline{X_M}}\}\}, & X_M - \overline{R_W} < 0 \\ \max\{p_w, \min\{\overline{p_0}, p_{m_W}, \frac{A_W a_W \overline{R_W}}{X_M}\}\}, & otherwise \end{cases} \tag{24}$$

where $\overline{p_0} = e^{-A_W a_V e^Y + \ln(A_V a_V) - 1} > 0$.

Proof. See technical report [12].

After $\overline{p}$, w, $\overline{m_W}$, and $\sum_{i=1}^{I} R_{F_i} m_{F_i}$ are determined, we can calculate the maximum value of $\overline{U_W}$. Comparing $\overline{U_W}^*$ with U_W^* obtained in the situation that the WSP serves the VMO by itself, the WSP can choose the more profitable way to serve the VMO.

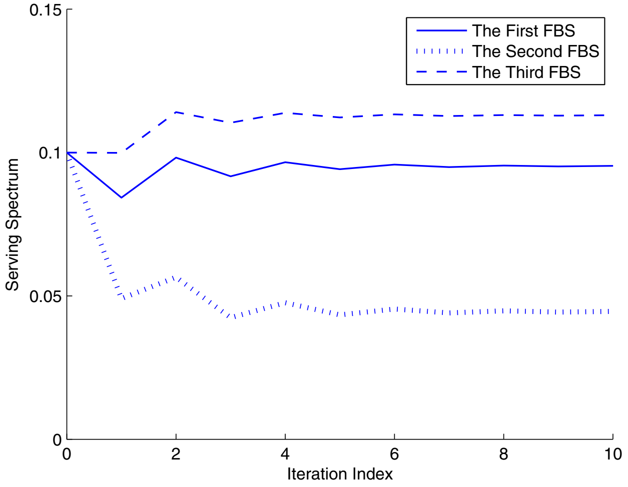


Fig. 2. Spectrum used to serve VUs versus iteration index

5 Simulation Results and Discussion

In this section, we conduct several simulations to study the behaviours of VMO, WSP and FBSs with the purpose to maximize the utility value of WSP.

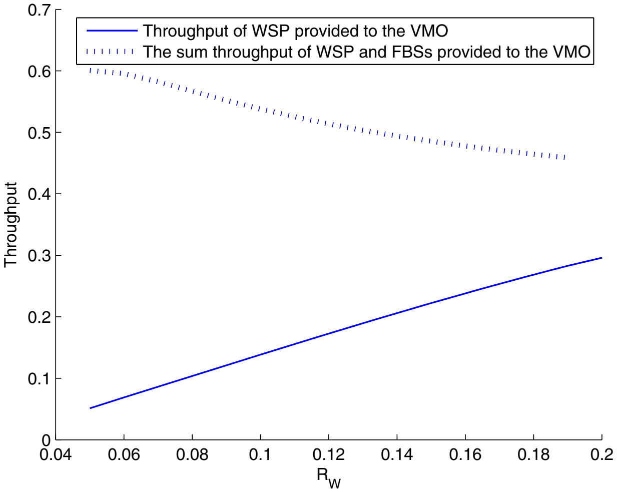


Fig. 3. Total throughput versus the data rate of WSP

Here are some simulation settings. We suppose that there are several users who want a specific VMO's services. Meanwhile they are located closely in the coverage of one WSP and three FBSs. In simulations, we assume the VUs will select the nearest MBS or FBS to access. For simplicity, the WSP is supposed to have the same data rate when it serves different users located in the same place. And so do FBSs. The inter-interference between the WSP and FBSs is eliminated because they use different spectrum. There is also no interference among FBSs since they are located sparsely. The adjustment coefficients of the satisfaction are $A_V = 2$ for the VMO, $A_W = 1$ for the WSP and $A_F = 1$ for FBSs. The coefficients weighting sensitivity of changed services are $a_V = 3$ for the VMO, $a_W = 50$ for the WSP and $a_F = 50$ for FBSs. Higher value of these factors means less sensitive towards throughput. The whole spectrum resources owned by the WSP are $W = 2MHz$.

Fig.2 shows the convergence of the FBSs' best response when they compete with each other at the fourth stage of the four-stage Stackelberg game. The data rate of three FBSs is supposed to be $R_{F1} = 1.2Mbps$, $R_{F2} = 1Mbps$ and $R_{F3} = 1.3Mbps$. The data rate of WSP is $R_W = 0.15Mbps$. The vertical axis represents the spectrum bands that every FBS splits to the VMO. In Fig. 2, three FBSs' strategies are supposed to be the same at the beginning and then converge after the eighth time slot. Clearly, different FBSs have different best response strategies because of the different data rates, which is supported by Theorem 3. The existence and uniqueness of Nash equilibrium in the NSCG are also verified.

In the following, in order to simplify the analysis and highlight the objective of femtocells, we assume the three FBSs have the same rate $R_{F_i} = 1Mbps$. Fig.3 shows the throughput of VMO versus the data rate of the WSP. When the WSP serves the VMO by itself, it provides less throughput if it has lower R_W. On the contrary, FBSs provide better services due to their high data rate. If the R_W becomes bigger, these two service ways will provide similar services undoubtedly. The service provided by FBSs interrupts near $R_W = 0.19Mbps$ because at this point the spectrum w leased to FBSs comes down to zero. The similar service interruption of FBSs can also be found in the next two figures.

Fig.4 shows the behaviour that the WSP adjusts its pricing strategy based on different R_W. When the WSP serves the VMO by its own, the price per unit throughput is very high at the beginning because the WSP wants to earn more although it can not provide large throughput. The price becomes lower with the increment of R_W. On the other hand, when the WSP selects to ask FBSs to help serve, the price per unit throughput is lower than the situation that the WSP serves the VMO by its own. However, because the WSP provides more throughput with FBSs' help, it can get a larger amount of revenue from the VMO.

The WSP's utility under two different situations are shown in the Fig.5. At the beginning, since R_W is very low, the utility value of WSP serving by its own is lower than that with FBSs. In this situation, the WSP prefers to ask FBSs to help serve obviously. However, when R_W becomes close to R_F, the WSP can provide services similar to FBSs'. That is to say, renting FBSs will waste more spectrum resources and earn less profits. It is found in Fig.5 that the utility value of WSP serving without FBSs comes to the same as serving with FBSs at $R_W = 0.13Mbps$. And it is wise for WSP to serve by its own when R_W becomes higher than $0.13Mbps$.

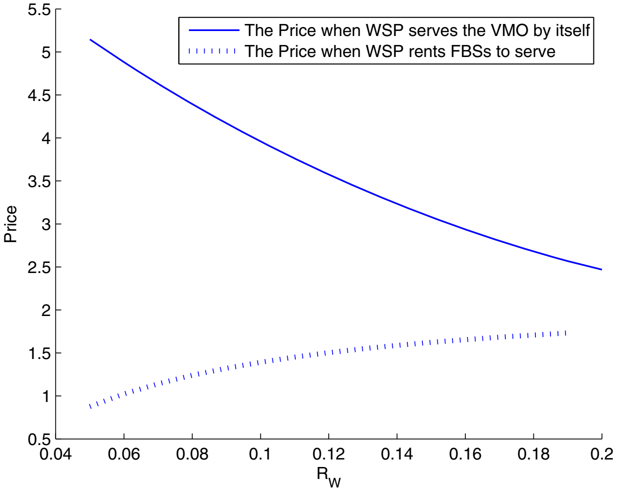


Fig. 4. Price versus WSP's data rate

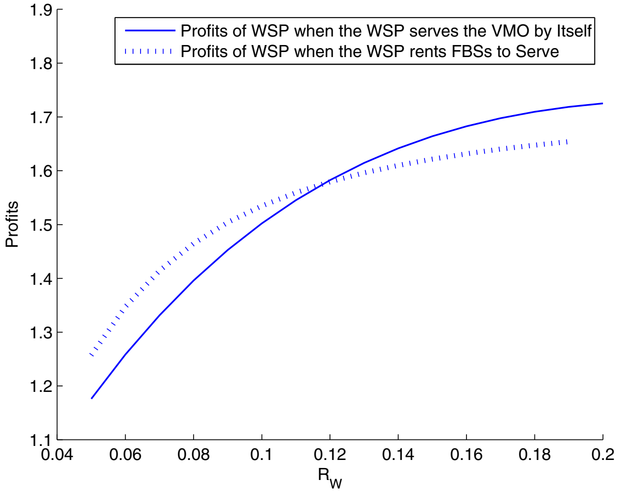


Fig. 5. Profits versus WSP's data rate

6 Conclusion

In this paper, we consider the problem that the VMO does not own spectrum resources and buy the wireless service from the WSP to serve its own users. The WSP, owning licensed spectrum and MSBs, is willing to serve VMO for extra income. A two-stage Stackelberg game is built to describe this scenario. In addition, WSP should ask FBSs to help serving VMO in poor coverage areas by offering FBSs free spectrum bands. We also build a four-stage Stackelberg game to analyze the second scenario. Both games are analyzed by backward induction method. After analyzing the competition among FBSs, we prove the existence and uniqueness of Nash equilibrium in the formulated non-cooperative game among FBSs. As to the two service mode, the WSP will choose one with higher profits, which corresponds to different coverage quality of WSP. In the simulation part, the existence and uniqueness of Nash equilibrium among FBSs is verified. And simulation results also shows that renting FBSs is better when the WSP faces poor coverage.

Acknowledgement. This work was supported partially by the NSFC under Grant 60934003, 61004062, 61174127, 61104033, 61273181, 61203104, 61221003, the National Basic Research Project of China under Grant 2010CB731800, the Research Found for the Doctoral Program of Higher Education under Grant 20110073120025, 20110073130005 and the Cyber Joint Innovation Center. This work was also supported partially by scientific research plan projects of Hebei Education Department under grant No.Q2012088 and Nature Science Foundation of Hebei Province F201220312.

References

1. Cisco VNI Mobile. Cisco visual networking index: Global mobile data traffic forecast update, 20122017. White Paper (February 2013)
2. Mansfield, G.: Femtocells in the us market-business drivers and consumer propositions. FemtoCells Europe, 1927–1948 (2008)
3. Neruda, M., Vrana, J., Bestak, R.: Femtocells in 3g mobile networks. In: 16th International Conference on Systems, Signals and Image Processing, IWSSIP 2009, pp. 1–4. IEEE (2009)
4. Duan, L., Huang, J., Shou, B.: Investment and pricing with spectrum uncertainty: a cognitive operator's perspective. IEEE Transactions on Mobile Computing 10(11), 1590–1604 (2011)
5. Xie, R., Yu, F.R., Ji, H.: Energy-efficient spectrum sharing and power allocation in cognitive radio femtocell networks. In: 2012 Proceedings of IEEE, INFOCOM, pp. 1665–1673. IEEE (2012)
6. Yi, Y., Zhang, J., Zhang, Q., Jiang, T.: Spectrum leasing to femto service provider with hybrid access. In: 2012 Proceedings of IEEE, INFOCOM, pp. 1215–1223. IEEE (2012)
7. Yun, S., Yi, Y., Cho, D.-H., Mo, J.: Open or close: on the sharing of femtocells. In: 2011 Proceedings of IEEE, INFOCOM, pp. 116–120. IEEE (2011)
8. Arulselvan, N., Ramachandran, V., Kalyanasundaram, S., Han, G.: Distributed power control mechanisms for hsdpa femtocells. In: 69th IEEE Vehicular Technology Conference, VTC Spring 2009, pp. 1–5. IEEE (2009)
9. Li, X., Qian, L., Kataria, D.: Downlink power control in co-channel macrocell femtocell overlay. In: 43rd Annual Conference on Information Sciences and Systems, CISS 2009, pp. 383–388. IEEE (2009)
10. Chen, Y., Zhang, J., Zhang, Q.: Utility-aware refunding framework for hybrid access femtocell network. IEEE Transactions on Wireless Communications 11(5), 1688–1697 (2012)
11. Khanafer, A., Saad, W., Basar, T., Debbah, M.: Competition in femtocell networks: Strategic access policies in the uplink. In: 2012 IEEE International Conference on Communications (ICC), pp. 5070–5074. IEEE (2012)
12. Yang, F., Yang, B., Guan, X.: A game theoretic service trading between the virtual mobile operator and the wireless service provider with assistance of femtocell. Tech. Rep., `http://wicnc.sjtu.edu.cn/techreport/yf.pdf`

Truthful Online Reverse Auction with Flexible Preemption for Access Permission Transaction in Macro-Femtocell Networks

Tao Jing[1], Fan Zhang[1], Liran Ma[2], Wei Li[3], Xuhao Chen[4], and Yan Huo[1]

[1] School of Electronics and Information Engineering, Beijing Jiaotong University, China
[2] Department of Computer Science, Texas Christian University, Fort Worth, TX, USA
[3] Department of Computer Science, The George Washington University, Washington, DC, USA
[4] Department of Computer and Information Science, Fordham University, New York, NY, USA

Abstract. In this paper, we study the problem of trading access permissions (ACPs) between a wireless service provider (WSP) and femtocell owners by truthful auctions. We propose a **T**ruthful **O**nline **R**everse **A**uction (TORA) mechanism that allows the WSP to purchase ACPs at a lower cost in an online manner while preventing femtocell owners from falsely reporting their bids and/or available time. To be specific, we develop an efficient allocation method with flexible preemption so as to enable a multiple-round online allocation and provide bid-truthfulness. In addition, we devise an effective pricing strategy so as to realize time-truthfulness. To the best of our knowledge, we are the first to study the truthful online reverse auction in a hybrid macro-femtocell network. We analytically prove the truthfulness of TORA. Our proof also shows that the truthfulness of TORA does not depend on the knowledge of the bidder behavior. An extensive evaluation study is performed to examine the performance of TORA. Our evaluation results indicate that TORA is able to achieve better bidder satisfaction with a lower cost.

Keywords: Macro-Femtocell Networks, Truthful Online Reverse Auction, Flexible Preemption, Access Permissions.

1 Introduction

There will be numerous wireless devices around us in the near future. It is predicted that 50 billion consumer electronics (such as smartphones) will be connected by 2020 [1]. These emerging devices are stressing the established macrocell networks beyond their original design objectives such as the coverage and capacity. Femtocell is a low-cost and low-power cellular base station that is designed to improve the indoor coverage and offload traffic from overburdened macrocell networks [2,13,17]. Hence, the integration of these two types of cellular networks becomes a cost-effective solution for a wireless service provider (WSP). The introduction of a hybrid macro-femtocell network entails many challenges. One fundamental challenge is how to fairly trade spectrum access permissions (ACPs) between a WSP and femtocell owners. To be specific, the WSP seeks to purchase a certain number of ACPs at various locations with a reasonable price so as to provide better services to its clients, while the femtocell owners compete to

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 512–523, 2013.

sell their idle ACPs to the WSP so as to maximize their revenue. This is challenging in a hybrid macro-femtocell network because the trade needs to be done truthfully in an online fashion.

Auction based mechanisms have been widely used for trading spectrum resources in wireless communication systems [5]. Since it is often the case that the number of the WSPs is far fewer than that of the femtocell owners in a hybrid macro-femtocell network, the reverse auction model [4] (i.e., an auction with multiple sellers and a single buyer) needs to be adopted instead of the conventional auction models [3, 8, 16, 19]. Most existing work on the auction mechanism design focuses on either a single-round auction or a periodic auction [3, 8, 16, 19]. The periodic auction refers to the scenario where the auctioneer periodically executes the auction procedure to perform resource re-allocation. Due to the dynamics of a hybrid macro-femtocell network, the number of femtocell owners and the availability of theirs ACPs vary from time to time and locations to locations. As a result, a multiple-round online auction [14, 15], where the auction decision can be triggered by the arrival/departure of buyers or sellers, needs to be adopted. In an online auction, winner preemption [9] needs to be allowed when a new bidder presents a better offer (e.g., a lower price). However, winner preemption may incur penalties such as a reduced number of ACP transactions or an increased number of wasted ACPs. Therefore, a flexible preemption scheme is needed so as to strive a balance between benefits and penalties.

Moreover, another critical feature of an auction mechanism is to guarantee the truthfulness. Otherwise, both the sellers and the buyers are reluctant to participate the auction because of the fear of market manipulation [11]. In a hybrid macro-femtocell network, a femtocell owner can cheat by falsely reporting either its bid and/or the availability (i.e., the start time and end time) of its ACPs. For example, a selfish femtocell owner can acquire a higher payment by strategically altering the bid and the reported start time of its ACPs to match the needs of the WSP. Hence, the truthfulness requirement in a hybrid macro-femtocell network is twofold: i) bid-truthfulness; and ii) time-truthfulness.

To address these challenges, we design a **T**ruthful **O**nline **R**everse **A**uction mechanism, termed **TORA**, for hybrid marco-femtocell networks. TORA allows the WSP to purchase ACPs at a lower cost in an online manner while preventing femtocell owners from cheating their bids and/or available time. To be specific, we develop an efficient allocation method with flexible preemption so as to enable a multiple-round online allocation. In addition, we devise an effective pricing strategy so as to achieve both bid- and time-truthfulness. We analytically prove the truthfulness of TORA. An extensive evaluation study is performed to evaluate the performance of TORA. Our evaluation results indicate that TORA is able to achieve better bidder satisfaction with a lower cost.

The rest of the paper is organized as follows. Preliminary is illustrated in Section 2. The design of our proposed auction mechanism TORA is described in Section 3. The detailed analysis of our auction mechanism appears in Section 4. The evaluation results are reported and analysed in Section 5. The related work is summarized in Section 6. We finally conclude our paper in Section 7.

2 Preliminary

In this section, we first give a detailed description of the auction scenario in our consideration. Next, we introduce the concept of truthful auction and its requirements.

2.1 Auction Scenario

We consider a time-slotted macro-femtocell network with one WSP and N femtocells as shown in Fig. 1. Each femtocell covers a certain area. The covered area can be further divided into several locations. At each location, a femtocell owner may have different number of ACPs available for sell in a given period of time. Assume that the WSP intends to purchase a number of ACPs in some specific locations. Each femtocell owner, referred to as a bidder, submits a bid to the WSP in a concealed fashion. The bid from a femtocell owner i contains the following information: i) The number of available ACPs (c_i) at each location; ii) The availability (i.e., the start time s_i and the end time e_i) of the available ACPs; iii) The bid price (b_i) for a unit ACP. The WSP needs to make ACP purchase decisions once the bid information is collected from the femtocell owners.

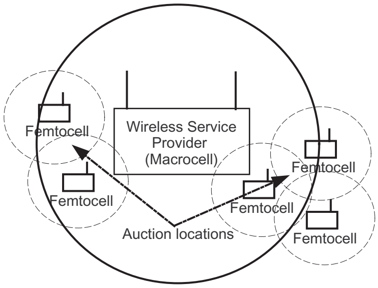


Fig. 1. The auction scenario for a hybrid macro-femtocell network

2.2 Truthfulness Conditions for an Online Reverse Auction

Truthfulness is one of the critical features required by an economic-robust auction [19]. In conventional auction theory, an auction mechanism is bid-truthful if it satisfies two conditions: *monotonicity* and *critical-value payment* [18]. The monotonicity means that if bidder i wins the auction by bidding b_i, bidder i can also win by bidding $b_i' \geqslant b_i$. The critical value means the lowest bidding value that a bidder can win. Hence, the critical value can be used as the price paid to the winner. This bid-independent price ensures that bidders cannot increase their utility gains by rigging their bids. Note that the critical value in a reverse online auction is defined as the highest bid that a bidder can submit in order to win the current time slot. Additionally, bidders may arrive randomly in time and confront different competitors in our online reverse auction. Therefore, the start time and the end time of the ACPs from a bidder also play an important role in the competition. We extend the truthfulness requirements of an online reverse auction to resist both bid and time cheating in Definition 1.

Definition 1. *An online reverse auction is (s, e, v) truthful if no bidder i can raise its utility gains by bidding $b_i \neq v_i$ or $[s'_i, e'_i) \subseteq [s_i, e_i)$.*

The tuple (s, e, v) refers to true start time, true end time, and true evaluation, respectively. Table 1 lists the notations and definitions used in our paper. To guarantee (s, e, v) truthfulness, it needs to satisfy two conditions: *monotonic allocation* and *temporal critical value based pricing*.

Table 1. Notations and Definitions

Symbol	Semantics	Symbol	Semantics
A_t	Auction event occurred at t	T	Minimum time units for using each ACP by WSP
s_i	The true start time of bidder i	e_i	The true end time of bidder i
s'_i	The reported start time of bidder i	e'_i	The reported end time of bidder i
b_i	The bid of bidder i	v_i	True evaluation of bidder i
p_{ij}	The price for each ACP payed to winner i at location j	p_{locj}	The payment of the WSP at location j
c_{ij}	The capacity of ACP of bidder i at location j	N_{ci}	The number of covered locations for bidder i
I	The preemption intensity	Req_j	The number of ACPs demanded by the WSP at location j
N_l	The total locations around the WSP	τ	The moment when a new bidder starts or a winner bidder finishes its sale
$\beta_i(t)$	The maximum value for bidder i to sell ACP(s) within continuous time period $[t, t+T-1]$	$\alpha_i(t)$	The critical value of bidder i at moment t

We give the definition of a monotonic allocation as follows:

Definition 2. *In an auction, if bidder i wins by bidding b_i and also wins by bidding $b'_i \leq b_i$, assuming that all other conditions remain the same, the resulting allocation is monotonic.*

In an online reverse auction, an allocation decision is made at a specific time τ, which is the moment that a new bidder arrives or a winner bidder finishes its sale. The monotonic allocation condition assures that there exists a critical value for bidder i in an auction A_τ. The WSP needs to compute the critical value (i.e., the highest bid i to win) of each time slot within a given time period. We use $\alpha_i(t)$ to present the critical value of bidder i at moment t. Subsequently, the maximum bid $\beta_i(t)$ that i can bid to win the T continuous time slots within $[t, t+T-1]$ for each $t \in [s'_i, e'_i - T]$ is defined as: $\beta_i(t) = \min\{\alpha_i(t) | t' \in [t, t+T-1]\}$. The temporal critical value of winner i can be calculated as follows:

$$p_i = \max_{t \in [s'_i, e'_i - T]} \beta_i(t). \tag{1}$$

The final payment is calculated based on the temporal critical value p_i.

3 Our Auction Mechanism

There are two core components in TORA: *Allocation* and *Pricing*.

3.1 Allocation

Since the ACP demand of the WSP and the number of available ACPs from a femtocell owner vary at different locations, we first illustrate a basic method to allocate ACPs at each location at a given time τ without preemption. The basic allocation method operates as follows:

1. For each location j at where the WSP demands ACPs, the WSP sorts the bids of the femtocell owners that cover the location j in a non-decreasing order to get the sorted list L^j and purchases ACPs from the top bidder $i \in L^j$. If the demand at location j is satisfied, the WSP goes on checking the next location (next iteration); otherwise, the WSP purchases ACPs from the next top bidder in the list L^j. This procedure continues until either the ACP demand of the WSP at location j is satisfied, or all the femtocell owners covering location j (L^j) run out of ACPs to sell.
2. At the end of each iteration, the remaining demand of the WSP at each location and the number of available ACPs of each femtocell owner are updated.

If a newly arrived bidder offers a better bid (e.g., a lower price), TORA may allow the current winner bidder to be preempted so as to reduce the cost. A preempted winner bidder cannot receive any payments if it supplies ACPs for fewer than T continuous time slots. However, a preempted winner bidder can be reconsidered for the next round auction before its end time. For a winning bidder i who has sold ACP(s) from moment t to moment t', TORA treats its bid as $\widetilde{b_i}(t')$ when ranking bidders at moment t':

$$\widetilde{b_i}(t') = b_i \cdot I^{\theta_i} \leq b_i, \tag{2}$$

where $\theta_i = (t' - t)/T$ represents the completion degree by time t, and I refers to the preemption intensity. Note that $I = 1$ means that if the bid of a newly arrived bidder is lower than that of the current existing winner i, i will be preempted. By decreasing I, the WSP adds more protection to the existing winners by ensuring a smaller probability of preemption. When $I \to 0, \widetilde{b_i}(t) \to 0$, the existing winner i will not be preempted. Algorithm 3.1 presents the detailed allocation procedure for a specific time τ. The notations are explained as follows: N_l is the total number of locations in the network; Req_j is the number of ACPs demanded by the WSP at location j; B_j is the set of bids where the bidders (femtocell owners) cover location j; $Continuous(i, j)$ is the estimated number of continuous time slots that the WSP has bought ACPs from bidder i before the current time at location j; $Low(B_j)$ refers to the bidder with the lowest bid in B_j; a_{qj} represents the quantity of the traded ACPs of bidder q at location j; c_{qj} indicates the number of available ACPs of bidder q at location j; if bidder q has sold ACPs at $\tau - 1$, but fails to sell ACPs at τ, we use $Preempt(q, \tau)$ to indicate that bidder q is preempted at time τ.

Algorithm 3.1: TORA Allocation Method

```
Require: i) the set of active bids $B$ at the current time $\tau$; and ii) the preemption intensity $I$.
Delete bidder $i$ if $e_i - s_i < T$;
for $j = 1$ to $N_l$ do
    if $Req_j > 0$ then
        $\widetilde{B_j} = \varnothing$.
        for $b_i \in B_j$ do
            $\theta_i = Continuous(i, j)/T$
            $\widetilde{b_i} = b_i \cdot I^{\theta_i}$
            $\widetilde{B_j} = \widetilde{B_j} \cup \widetilde{b_i}$
        end for
        while $\widetilde{B_j} \neq \emptyset$ do
            $q = Low(\widetilde{B_j})$
            if $Req_j > 0$ then
                $a_{qj} = \min\{Req_j, c_{qj}\}$
                $Req_j = Req_j - a_{qj}$
            else
                if $q$ sold ACPs at $(\tau - 1)$ then
                    $Preempt(q, \tau)$
                end if
            end if
            $\widetilde{B_j} = \widetilde{B_j} \setminus \{\widetilde{b_q}\}$
        end while
    end if
end for
```

3.2 Pricing

The pricing stage determines payments for winner femtocell owners. Assume that the available time period of winner i is $[s_i, e_i)$. The payment p_i for winner i is determined by the *interval payment* $\beta_i(t)$, where $t \in [s_i, e_i - T]$. The interval payment $\beta_i(t)$ is defined as the maximum value that i can bid to win T continuous time slots starting from moment t. The value of $\beta_i(t)$ depends on the maximum bid $\alpha_i(t')$ that is required for i to win (i.e., the critical value for bidder i) for each moment $t' \in [t, t + T - 1]$. To be specific, the interval payment is calculated as follows:

$$\beta_i(t) = \min_{t' \in [t, t+T-1]} \frac{\alpha_i(t')}{I^{(t'-t)/T}}, \tag{3}$$

where I is the preemption intensity, and $(t' - t)/T$ represents the completion degree by time t. Note that the approach of calculating the critical value $\alpha_i(t')$ is adopted from [18]. After calculating all possible interval payments for each $t \in [s_i, e_i - T]$, the WSP pays i the maximum interval payment. Accordingly, the payment of winner i for each ACP at e_i can be computed as:

$$p_i = \max_{t \in [s_i, e_i - T]} \{\beta_i(t)\}. \tag{4}$$

Algorithm 3.2 presents the temporal critical value based pricing procedure in time frame $[0, Total]$. It details the procedure of computing the payment p_i for each ACP of bidder i at its end time and the total amount that the WSP needs to pay for the transacted ACPs. The notations and functions are described as follows: $Fir(\mathcal{L}_j)$ refers to the first bidder in the sorted list $\mathcal{L}_j$, where $\mathcal{L}_j$ is the sorted list of the femtocell owners that cover location j in non-decreasing order of their start time; $timelength(i, j)$ means the longest continuous time slots for selling ACPs of bidder i at location j; $Ongoingbid(B, j, t)$ indicates the set of bidders whose start time is no later than t and end time is later than t; $CriticalValue(\widetilde{B_j}, i, t, j)$ returns the critical value of bidder i on the current time slot t at location j; p_{locj} returns the payment of the WSP at location j; P_{wsp} indicates the total payment that the WSP needs to pay for all locations.

Algorithm 3.2: TORA Pricing

```
Require: i) the set of bids $B$ during the time range $[0, Total]$; and ii) the preemption
    intensity $I$.
for $j = 1$ to $N_l$ do
  if $Req_j > 0$ then
    while $\mathcal{L}_j \neq \varnothing$ do
      $i = Fir(\mathcal{L}_j)$
      if $timelength(i, j) < T$ then
        $p_{ij} = 0$
      else
        for $t \in [s_i, e_i - 1]$ do
          $\widetilde{B_j} = \varnothing$
          $list = Ongoingbid(B, j, t)$
          for $x \in (list \backslash \{i\})$ do
            $\theta_x = Continuous(x, j)/T$
            $\widetilde{b_x} = b_x \cdot I^{\theta_x}$
            $\widetilde{B_j} = \widetilde{B_j} \cup \widetilde{b_x}$
          end for
          $\alpha_i(t) = CriticalValue(\widetilde{B_j}, i, t, j)$
        end for
        for $t \in [s_i, e_i - T]$ do
          $\beta_i(t) = \min\{\frac{\alpha_i(t')}{I^{(t'-t)/T}} \mid t' \in [t, t + T - 1]\}$
        end for
        $p_{ij} = \max\{\beta_i(t) \mid t \in [s_i, e_i - T]\}$
      end if
      $\mathcal{L}_j = \mathcal{L}_j \setminus \{i\}$
    end while
    $p_{locj} = \sum_{i \in \mathcal{L}_j} a_{ij} \cdot p_{ij}$
  end if
end for
$P_{wsp} = \sum_{j \in N_l} p_{locj}$
```

4 Theoretical Analysis

In this section, we prove that TORA achieves both bid- and time-truthfulness (i.e., (s, e, v)-truthful). That is, bidders cannot improve their utility by bidding $b_i \neq v_i$ or $[s'_i, e'_i) \subseteq [s_i, e_i)$. Let u_i be the utility with start time, end time, and valuation being (s_i, e_i, v_i), and u'_i be the utility with start time, end time, and bid being (s'_i, e'_i, b_i). Denote p_i and p'_i as the payment that bidder i gets at each case, respectively.

We start by proving that Algorithm 3.1 satisfies *monotonicity* with the inclusion of the preemption factor I.

Lemma 1. *Algorithm 3.1 is a monotonic allocation.*

Proof. As described in Algorithm 3.1, the WSP sorts all the bids that cover the same location in a non-decreasing order. Subsequently, the WSP allocates ACP purchases following this order. We use $r(b_i)$ and $r(b'_i)$ to represent the rank of bidder i when i bids with b_i and b'_i, respectively. If bidder i bids $b'_i \leq b_i$, we have $b'_i \cdot I^{\theta_i} \leq b_i \cdot I^{\theta_i}$. Hence, we can get that $r(b'_i) \leq r(b_i)$. If i has sold ACPs at τ by bidding b_i, i can also sell ACPs when bids $b'_i \leq b_i$. This completes our proof.

Lemma 2. *Assume that bidder i wins by bidding either (s_i, e_i) or (s'_i, e'_i), where $s_i \leq s'_i$ and $e_i \geq e'_i$. Let p_i and p'_i represent the payment for (s_i, e_i) and (s'_i, e'_i), respectively. We have $p_i \geq p'_i$.*

Proof. When $t \in [s'_i, e'_i - T]$, according to Eq. (3), we can get that $\beta_i(t)' = \beta_i(t)$. As $[s'_i, e'_i - T] \subseteq [s_i, e_i - T]$, thus we can derive the following inequality: $p'_i = \max_{t \in [s'_i, e'_i - T]}\{\beta_i(t)'\} \leq \max_{t \in [s_i, e_i - T]}\{\beta_i(t)\} = p_i$. This completes our proof.

Theorem 1. *TORA is (s, e, v)-truthful.*

Proof. We show that $u_i \geq u'_i$ for all cases as follows.

1. i loses with both bids (s_i, e_i, v_i) and (s'_i, e'_i, b_i): $u_i = u'_i = 0$.
2. i loses with (s_i, e_i, v_i) and wins with (s'_i, e'_i, b_i). According to the monotonicity, it happens only when $b_i \leq v_i$. Since $p'_i = \max\{\beta_i(t) \mid t \in [s'_i, e'_i - T]\} \leq \max\{\beta_i(t) \mid t \in [s_i, e_i - T]\} \leq v_i$, we can conclude that $p'_i \leq v_i$. So $u'_i = p'_i - v_i \leq u_i = 0$.
3. i wins with (s_i, e_i, v_i) and loses with (s'_i, e'_i, b_i). Assume starting from t, $t \in [s_i, e_i - T]$, bidder i wins at least T continuous time slots. For each $t' \in [t, t + T - 1]$, $\alpha_i(t') \geq v_i \cdot I^{(t'-t)/T}$. As a result, $\beta_i(t) = \min\{\frac{\alpha_i(t')}{I^{(t'-t)/T}} \mid t' \in [t, t + T - 1]\} \geq v_i$. Thus, $p_i = \max\{\beta_i(t) \mid t \in [s_i, e_i - T]\} \geq v_i$. Hence, $u_i = p_i - v_i \geq u'_i = 0$.
4. i wins with either (s_i, e_i, v_i) or (s'_i, e'_i, b_i). As p_i and p'_i are independent with the bid value and $[s'_i, e'_i) \subseteq [s_i, e_i)$, we can infer that $p_i \geq p'_i$ according to Lemma 2. Therefore, $u_i = p_i - v_i \geq p'_i - v_i = u'_i$.

By showing that bidder i cannot get a higher utility gain by cheating in time and/or bid, we prove that TORA is (s, e, v)-truthful.

5 Evaluation

In this section, we investigate the performance of our proposed TORA through a simulation study. For comparison purposes, we choose the reverse auction scheme proposed in [4] as the reference scheme. Note that there does not exist a scheme that is commensurate with our work since we are the first to study the truthful online reverse auction in a hybrid macro-femtocell network.

5.1 Settings

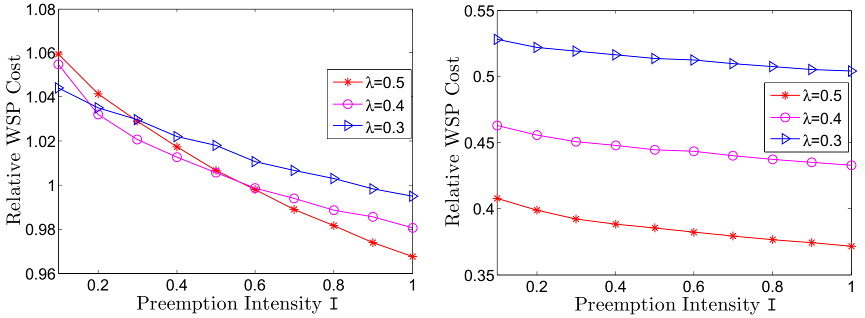
We assume that there are total N_l locations in the network. To be compatible with the mainstream research [14], the arrival of bidders is assumed to be following the Poisson distribution with an arrival rate of $\lambda \in (0, 1)$. For each femtocell owner i, we assume that i covers N_{ci} locations (randomly selected from N_l locations), and provides c_i number of ACPs for sell at each location. The value of c_i is uniformly distributed in $[4, 10]$. At each location that i covers, the value of c_i is generated independently. In addition, the availability of its ACPs, denoted by $e_i - s_i$, and the bid b_i, are uniformly distributed in $[5, 10]$ and $(0, 1]$, respectively. In some simulation scenarios, a femtocell owner can cheat by manipulating its bid. We assume that a cheating bidder may manipulate its bid value up to 5 times of the bid's true valuation. Since the auction scheme in [4] does not handle time cheating, we choose not to allow femtocell owners falsely report their start and/or end time. For the WSP, we assume that the WSP intends to purchase ACPs at N_p locations. At each of these location, the WSP needs to buy Req number of ACPs. To ensure the communication quality, the WSP requires each purchased ACP to be available for at least T time units. The major simulation parameters are listed in Table 2.

Table 2. Simulation Parameters

Symbol	Value	Symbol	Value
N_l	8	b_i	(0,1]
N_{ci}	4	$e_i - s_i$	[5,10]
Req	7	c_i	[4,10]
N_p	4	T	7

5.2 Results

We investigate the benefits of our proposed TORA in terms of the cost of the WSP and bidder satisfaction. Each simulation run lasts 500 unit time slots. The simulation results are averaged over 50 independent runs. Fig. 2 reports the relative cost of the WSP between the two schemes with the preemption index I varying from 0.1 to 1. When there are no cheaters present in the network, TORA may incur more cost from the WSP compare to that of the scheme proposed in [4] as shown in Fig. 2(a). However, where there are cheaters present in the network, TORA incurs much less cost from the WSP

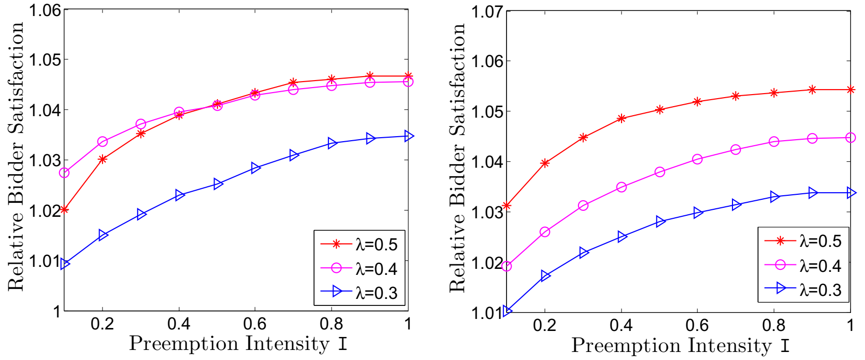


(a) Without cheating bidders

(b) With cheating bidders

Fig. 2. WSP cost

compare to that of the scheme proposed in [4] as shown in Fig. 2(b). The maximum cost reduction is about 62% achieved when I approaches 1 and $\lambda = 0.5$. Additionally, the cost of the WSP in TORA deceases with I.



(a) Without cheating bidders

(b) With cheating bidders

Fig. 3. Bidder satisfaction

Fig. 3 plots the relative bidder satisfaction between the two schemes under the identical settings as those of the WSP cost. The bidder satisfaction is defined as the sum of all winning bidders in each location in the network. Intuitively, a bidder exhibits a certain “satisfaction” if it gains a payment for successfully selling its ACPs at a location. Hence, the overall bidder satisfaction should increase with the total number of winners in each location in the network. We observe that TORA outperforms the scheme proposed in [4] in terms of bidder satisfaction for almost all cases. It also can be seen that the bidder satisfaction gently increases with I, which indicates that the total number of winners in each location in the network slightly increases with I as well.

6 Related Work

The existing mainstream work on the auction mechanism design focuses on either a single-round auction or a periodic auction [3, 8, 16, 19]. The periodic auction refers to the scenario that the auctioneer periodically executes the auction procedure to perform resource re-allocation. Due to the dynamics of a wireless network, the availability of the sellers and/or buyers varies from time to time and locations to locations. Hence, the single-round auction or periodic auction mechanisms perform poorly under such dynamics. To address this problem, the idea of online auction, where the auction decision can be triggered by the arrival/departure of sellers and/or buyers, is proposed in [9, 14, 15]. For instance, an online auction mechanism is proposed in [15] where two protocols are developed so as provide guaranteed performances for two different online scenarios. But the preemption is not allowed. Different from [9, 14, 15], our online auction scheme TORA allows flexible preemption, which is inspired by the scheme developed in [6] where it proposes to control the preemption intensity for performance enhancement. Femtocell ACP transaction has been considered in [4], which provides a sub-optimal efficient auction mechanism to reduce the computation complexity. But the auction in [4] is a periodic one. In our paper, we attempt to allocate ACPs of femtocells to the WSP in an online manner. Note that existing online auction schemes may not be directly applied to the reverse auction scenario.

There exist two pricing models in auction theory [7], i.e., *uniform pricing* and *discriminatory pricing*. Uniform pricing is a popular approach [19]. However, as bidders come randomly and need to be paid at the end time, it is intuitive for TORA to adopt the discriminatory pricing. There are two types of bid requests: i) *strict request*; and ii) *range request*. In [10], periodic truthful double auction is considered for strict request. However TORA accepts the range request from both the WSP and the femtocell owners.

When designing an auction mechanism, the truthfulness must be satisfied so as to attract the participation from both the sellers and the buyers [11]. Most existing work adopts the McAfee mechanism proposed in [12] to ensure the truthfulness. However, the McAfee mechanism only suits for the single-round auctions, and hence, cannot be applied to the online auction. Additionally, a general framework of truthful double online auction is proposed in [14]. The bid, time, and job truthfulness of online scheduling mechanism is considered in [9]. But the solutions in [9, 14] do not suit for the reverse online auction scenario. The work closest to ours is [6] where an online auction scheme based on 3D bin packing is developed to resist bid- and time-cheating for efficient spectrum redistribution among the competing users [6]. Nonetheless, the proposed 3D bin packing based scheme cannot be applied to our network scenarios where the ACP demands and supplies may vary at different locations.

7 Conclusion

In this paper, we propose a truthful online reverse auction mechanism termed TORA for femtocell owners and a WSP to trade ACPs. Inside TORA, an efficient allocation method with flexible preemption is developed so as to enable a multiple-round online allocation. In addition, we devise an effective pricing strategy so as to achieve both bid- and time-truthfulness. We analytically prove the truthfulness of TORA. An extensive evaluation study is performed to evaluate the performance of TORA. Our evaluation results indicate that TORA enables the WSP to purchase ACPs at a lower cost and improves the bidder satisfaction. As a part of our future work, we plan to consider a multi-attribute auction that considers various wireless link characteristics.

References

1. More than 50 billion connected devices. White paper, Ericsson, `http://www.ericsson.com/res/docs/whitepapers/wp-50-billions.pdf`
2. Chandrasekhar, V., Andrews, J., Gatherer, A.: Femtocell networks: a survey. IEEE Communications Magazine 46(9), 59–67 (2008)
3. Chen, L., Iellamo, S., Coupechoux, M., Godlewski, P.: An auction framework for spectrum allocation with interference constraint in cognitive radio networks. In: IEEE INFOCOM, pp. 1–9 (March 2010)
4. Chen, Y., Zhang, J., Zhang, Q., Jia, J.: A reverse auction framework for access permission transaction to promote hybrid access in femtocell network. In: IEEE INFOCOM, pp. 2761–2765 (March 2012)
5. Cramton, P.: Spectrum auctions. Papers of Peter Cramton 01hte, University of Maryland, Department of Economics - Peter Cramton (2002)
6. Deek, L., Zhou, X., Almeroth, K., Zheng, H.: To preempt or not: Tackling bid and time-based cheating in online spectrum auctions. In: IEEE INFOCOM, pp. 2219–2227 (April 2011)
7. Gandhi, S., Buragohain, C., Cao, L., Zheng, H., Suri, S.: A general framework for wireless spectrum auctions. In: DySPAN 2007, pp. 22–33 (April 2007)
8. Gopinathan, A., Li, Z.: Strategyproof auctions for balancing social welfare and fairness in secondary spectrum markets. In: IEEE INFOCOM, pp. 3020–3028 (April 2011)
9. Hajiaghayi, M.T.: Online auctions with re-usable goods. In: Proceedings of the 6th ACM Conference on Electronic Commerce, pp. 165–174 (2005)
10. Jia, J., Zhang, Q., Zhang, Q., Liu, M.: Revenue generation for truthful spectrum auction in dynamic spectrum access. In: MobiHoc 2009, pp. 3–12 (2009)
11. Klemperer, P.: What really matters in auction design. Journal of Economic Perspectives 16(1), 169–189 (2002)
12. McAfee, R.P.: A dominant strategy double auction. Journal of Economic Theory 56(2), 434–450 (1992)
13. de la Roche, G., Valcarce, A., Lopez-Perez, D., Zhang, J.: Access control mechanisms for femtocells. IEEE Communications Magazine 48(1), 33–39 (2010)
14. Wang, S., Xu, P., Xu, X., Tang, S., Li, X., Liu, X.: Toda: Truthful online double auction for spectrum allocation in wireless networks. In: IEEE Symposium on New Frontiers in Dynamic Spectrum, pp. 1–10 (April 2010)
15. Xu, P., Wang, S., Li, X.Y.: Salsa: Strategyproof online spectrum admissions for wireless networks. IEEE Transactions on Computers 59, 1691–1702 (2010)
16. Xu, W., Wang, J.: Double auction based spectrum sharing for wireless operators. In: PIMRC 2010, pp. 2650–2654 (September 2010)
17. Zhang, J., de la Roche, G.: Femtocells: Technologies and Deployment. Wiley Publishing (2010)
18. Zhou, X., Gandhi, S., Suri, S., Zheng, H.: ebay in the sky: strategy-proof wireless spectrum auctions. In: MobiCom 2008, pp. 2–13 (2008)
19. Zhou, X., Zheng, H.: Trust: A general framework for truthful double spectrum auctions. In: IEEE INFOCOM, pp. 999–1007 (April 2009)

Social Communications Assisted Epidemic Disease Influence Minimization

Bowu Zhang[1], Pei Li[2], Xiuzhen Cheng[1], Rongfang Bie[3], and Dechang Chen[4]

[1] Computer Science, The George Washington University, DC, USA
[2] College of Information Systems and Management,
National University of Defense Technology, ChangSha, China
[3] Information Science and Technology, Beijing Normal University, Beijing, China
[4] Division of Epidemiology and Biostatistics,
Uniformed Services University of the Health Sciences, MD, USA
{bowuzh,cheng}@gwu.edu, pli.nudt@gmail.com,
bierf@163.com, dechang.chen@usuhs.edu

Abstract. This work explores the use of social communications for epidemic disease control. Since the most infectious diseases spread through human contacts, we focus on modeling the diffusion of diseases by analyzing the social relationship among individuals. In other words, we try to capture the interaction pattern among human beings using the social contact information, and investigate its impact on the spread of diseases. Particularly, we investigate the problem of minimizing the expected number of infected persons by treating a small fraction of the population with vaccines. We prove that this problem is NP-hard, and propose an approximate algorithm representing a preventive disease control strategy based on the social patterns. Simulation results confirm the superiority of our strategy over existing ones.

Keywords: Preventive disease control, social networks, target vaccination.

1 Introduction

In this paper, we propose to use communication records to guide the use of vaccines in a population to minimize the impact of epidemic disease. Infectious diseases pose major risks to the human life and social development. In recent years, various infectious diseases such as H1N1 and SARS have caused thousands of deaths and severe economic loss. Most of the infectious diseases can be transmitted from one person to another through personal contacts. With rapidly growing global transportations, infectious diseases that formerly die out in isolated areas may now spread worldwide. In order to stop infectious diseases from spreading, a number of interventions are available with distinct benefits or drawbacks. They either directly impact the transmission of diseases so that the viruses/germs cannot easily spread through the population, or immunize segments of a population. Considering the time that interventions are applied,

K. Ren et al. (Eds.): WASA 2013, LNCS 7992, pp. 524–534, 2013.

strategies for controlling epidemic diseases can be classified into two categories: 1) preventive disease control that takes place before a disease occurs in a population; and 2) reactive disease control that stops a disease from spreading out after the disease outbreak is detected in a population. Reactive strategies focus on individuals who are already infected and their close friends to protect healthy people from being infected. Preventive strategies, on the other hand, identify people at high infection risks to take vaccines, so that the disease can be prevented from spreading out or even happening in a population.

In this paper, we design a preventive disease control strategy that takes action to set up defenses against disease breakout ahead of time. In particular, we attempt to immunize a small number of people who are in danger in advance, so that we can prevent them from developing deadly infections and spreading to others. Such preventive vaccination strategies are widely used to prevent diseases, i.e., people are encouraged to take flu shot before the flu season starts. However, current preventive disease control mainly relies on mass vaccination strategy, which intends to immunize a major fraction of a population, leading to a high economic expense and adverse side effects. Moreover, due to production cycle and population growth, vaccines are often in short supply, making the mass vaccination ineligible for a number of occasions. In light of the problems above, we aim to select individuals to receive vaccines in advance based on information extracted from people's daily life data stream, so that the number of people infected will be minimized if the population is exposed to a disease later.

Prior work that selects target individuals for vaccination mainly fall under two categories: 1) node centrality methods, which rank nodes by measures such as degree, shortest path, or random walk betweenness; and 2) influence cascade methods, which select a subset of nodes that maximize information diffusion using independent cascade models or linear threshold models. The latter has been widely adopted in many recent works due to the rising of the idea of viral marketing, where commercial messages are sent to people who are socially important in order to achieve marketing objectives. In this paper, on the contrary, we try to minimize the diffusion of disease, rather than maximizing the spreading of a piece of information. In particular, we show that our minimization problem is NP-hard, and can be solved by an approximation algorithm following the information cascade model. Our approach is evaluated over real world data set. The results demonstrate that the proposed target immunization outperforms other strategies in providing protection over the population.

Our contributions can be summarized as follows:

- We explore social communications to determine the pattern of disease transmissions among individuals.
- We attempt to minimize the expected number of infected individuals by treating a small fraction of the population. This problem is proved to be NP-hard in this paper and we propose an approximation algorithm that provides a simple preventive disease control strategy.
- A comparison based simulation study over a real-world data set is conducted to evaluate the performance of our preventive disease control strategy and

the results indicate that the proposed strategy is superior over the most popular existing ones.

The paper is organized as follows. Section 2 presents the most related work. Sections 3 discusses the preliminary knowledge to be used in this work. An important metric, the transition probability, is defined and analyzed in section 4. The disease control minimization problem and an approximation solution are detailed in section 5. The proposed strategy is evaluated under various scenarios in section 6. We conclude our work in section 7.

2 Related Work

Our work is built on considerable prior research on disease propagation and social relationship identification. In this section, we review the most related work along these directions.

Disease propagation has been studied for a long time in human contact networks [1–3], where human beings are interpreted as nodes and their interactions are edges connecting nodes. Various mathematical models have been developed to characterize the disease transmission in a contact network. Most of the existing models [4],[5] describe the spread of disease in a homogeneously mixed network, where an infected individual infects each of his/her neighbors in a uniformly random and independent way. Under the constraint of fixed network size, differential equations can be written down to represent the movement of disease in a network, from which the number of infected people can be estimated. However, in real world, an infected person does not infect his contacts with an equal probability, as the length and the nature of the interactions among people can vary greatly from one to another. In recent years, a few investigations have been made to study the impact of variations in degree, infectiousness, and closeness of interactions. For instance, different network topologies such as sparse networks [6], clustered networks [2], and power-law networks [7], have been examined for their effect on disease propagation process. In this research, we construct a contact network based on social communications, and model the spread of disease as a function of the social relationship between individuals.

Many works have offered practical insights into the impact of social relationship on the development of applications in various domains, ranging from monitoring/tracking applications, to medical, emergency, and military applications [8–10]. Among them, a large body of research has addressed the problem of viral marketing [11, 12], where the social connections are used to spread the product information to achieve business objectives. The work by [13] studies the information propagation in social networks using two information diffusion models: independent cascade (IC) and linear threshold. The paper proves the NP-hardness of the problem of finding a small number of influential nodes as initial information adopters to maximize the expected number of nodes that would adopt the information. In contrast to influence maximization (i.e., recent research on viral marketing mentioned above) that attracts attention from finance and social network communities, influence minimization has relatively been less investigated. Misinformation such as scams, false twitter or facebook posts, have led to enormous social and economic issues. In order to minimize the impact of misinformation, researchers have considered to block critical nodes to prevent the false message from spreading out. Such an idea has been applied in a number of contexts. In this work, we consider a similar minimization problem which we prove it to be NP-hard.

3 Preliminary

3.1 Contact Network

We investigate the spread of diseases over a contact network constructed over social communication records. Define the contact network as $G(V, E, W)$, where V denotes the node set, E denotes the edge set, and W represents the edge weight set. A node $v \in V$ represents an individual, while an edge e_{ij} between nodes i and j indicates that there exist interactions between the two nodes. Attached to the edge e_{ij} is the weight w_{ij}, which is used to quantify the interaction between i and j. If the interaction is made through phone calls, then w_{ij} can be the total time used in the phone conversations. If the interaction is done through messages, then w_{ij} can include information such as the number of the messages and the sizes of the messages.

3.2 The SIR Model

We use a SIR model to describe the progress of a disease in a single node. Nodes may be either *Susceptible* to the disease, or *Infected* with the disease, or *Recovered* from the disease. Assume an infective person stays *Infected* for α time intervals. During each time interval, an *Infected* node i infects a *Susceptible* node j that has interactions with i with a probability p. After α time intervals, an *Infected* node becomes *Recovered.* A *Recovered* node can not get infected or infect others. Nodes that receive vaccinations become *Recovered* automatically. Notice that we differentiate p from the transition probability defined in section 4, where p stands for the probability that a disease is transmitted between two nodes during one time interaction, while the transition probability states in general the probability for a node to pass a disease to another node. For the sake of simplicity, we assume p is the same for any two nodes. However, the transition probability, that we will discuss later, relies on the variations in closeness of human connections.

3.3 IC Model

We employ an independent cascade (IC) model [13] to study the spread of disease among people. In this model, the diffusion proceeds in discrete time intervals. Nodes in this model are either active or inactive. The diffusion starts with an initial active node set. Assume node i becomes active at step t. Then i will attempt to activate each of its inactive neighbors, j, with a transition probability

t_{ij}, which indicates the tendency of j to be activated by i. If i succeeds, j becomes active at step $t+1$. If j has multiple active neighbors at step t, their activation contacts with j would be sequenced in a random order. The IC process terminates if no more activations are possible. Denote by $\sigma(i)$ the *influence degree* of a node i, which is defined to be the expected number of nodes influenced by i at the end of the IC process. In this work, we aim to minimize the expected number of infected people by vaccinating a small fraction of the population.

4 Transition Probability

Since infectious diseases spread through human contacts, the disease transition probability strongly depends on the pattern of contacts between infected persons and others. Many works assume that this probability is the same between any two individuals, while in fact, it is obvious an infected person will not infect all others with the same probability. For example, the probability of a disease transmitted between two close friends is higher than the probability of disease transmission between two strangers. A number of factors influence the transition probability, including the vaccination history, the general health of the normal person, the nature of the disease, and the nature of the interaction (e.g., time, interaction type) between the individual and the infected person(s). In this work, we try to capture these factors based on social communication records so that we can predict how diseases are transmitted in a population.

Let T be a $N \times N$ matrix. Denote by t_{ij} the element of T at row i and column j, which states the transition probability describing how likely an epidemic disease will be passed from node i to node j. Considering the IC process which determines the diffusion of a disease, during each step, a disease only moves from the current node to its immediate neighbors. Therefore, for a pair of nodes that are not connected by edges, the transition probability between them is zero. For a pair of connected nodes, it is widely accepted that the transition probability between them depends on their interaction patterns [3, 14], since epidemic diseases spread through human contacts. In this work, we model the transition probability among individuals by analyzing the social relationship information in G over their contact records. Following our preliminary work [15, 16], we define the transition probability t_{ij} as follows:

$$t_{ij} = f(d_{ij}, N_i, N_j, w_{ij}, \sum_{k \in V, k \neq i} w_{ik}, \sum_{k \in V, k \neq j} w_{jk}) \tag{1}$$

where f is a multivariate function ranging from 0 to 1, N_i represents the set of nodes which communicate with v_i directly, and d_{ij} denotes the physical distance between i and j, which can be roughly estimated through cell phone's built-in GPS, or through cell phone tower log. Note that two individuals who have a large number of phone interactions may not have physical interactions through which epidemic disease can be transmitted. Therefore, we include d_{ij} to ensure that t_{ij} in the above case is not high. In addition, we let $t_{ii} = 0$ to avoid self-loop. The computation of t_{ij} can be completed locally at i and j, without requiring the global network information, which leads to low communication overhead. Further discussions on t_{ij} can be found in our preliminary work [15, 16].

5 Disease Influence Minimization

Different from target selection in viral marketing, which attempts to maximize the number of nodes that can be affected by the target set, we try to minimize the impact of disease over G by immunizing a small number of nodes. Denote by K a pre-defined constant satisfying $|K| < N$. Given a network $G(V, E, W)$ and the transition matrix T, we aim to immune a set A of K nodes, $A \subset V$, so that the number of nodes infected will be minimized if disease occurs. Assume that disease appears at each node with a probability q, where q is a positive constant less than 1. Thus after vaccinating nodes in A, the expected number of nodes infected by $v, v \in A^-$, is $q \cdot \sigma(v)_{\text{over } G^{A^-}}$, where A^- is the complementary set of A, and G^{A^-} is the subgraph of G induced by all the nodes in A^-. Therefore, the sum of the expected number of nodes infected by each node in A^-, denoted by $AVG(A^-)$, is $\sum_{v \in A^-} q \cdot \sigma(v)_{\text{over } G^{A^-}}$. Then our goal is to find A of K nodes for vaccination, so that $AVG(A^-)$ is minimized. Notice here $AVG(A^-)$ measures the sum of the expected number of nodes infected by each node in A^-, not the sum of the union of the expected number of nodes infected, in case that the nodes infected/influenced by $i \in A^-$ and the nodes infected/influenced by $j \in A^-$ overlap. The definition of $AVG(A^-)$ is based on the consideration that we attempt to minimize the impact of disease no matter which node the disease starts with. Then the problem can be mathematically described as:

$$\min_{A \subseteq V} AVG(A^-) \text{ s.t } |A| = K \tag{2}$$

To solve (2), intuitively, we can check every possible set of K nodes in V, which takes $\binom{N}{K} = O(n^K)$ time. We will prove in the following section that this minimization problem is NP-hard, which can not be solved efficiently. An approximation algorithm is then proposed to provide an approximate solution.

5.1 NP-hardness

Consider the following sum-of-squares partition problem [17]. Let $G(V, E)$ be an undirected graph, with a node set V and an edge set E. Given a constant K, the sum-of-squares problem attempts to partition G into disconnected components $C_1, \ldots, C_i, \ldots$ by removing a set A of at most K nodes such that $\sum_i |C_i|^2$ is minimized. This problem is known to be NP-hard. We present in the following, that we can reduce the sum-of-squares partition problem to the proposed minimization problem in polynomial time when the transition probability is 1.

Theorem 1. *When the transition probability is 1, finding a node set $A \subseteq G, |A| = K$ so that $AVG(A^-) \leq AVG(\Theta^-)$ for $\forall \Theta \subseteq V, |\Theta| = K$ is a NP-hard problem.*

Proof. As the transition probability between any two nodes in G is 1, once $v \in G$ is infected, all the nodes that can be reached from v will be infected. Then for any node v in C_i, if v is infected, the expected number of influenced nodes is $|C_i|$. Thereby the sum of the expected number of infected nodes is $AVG(A^-) = \sum_i q \cdot |C_i|^2$. As a result, minimizing $\sum_i |C_i|^2$ is equivalent to minimizing $AVG(A^-)$ in G. Through the above steps, we successfully transform the sum-of-squares partition problem to the problem of minimizing $AVG(A^-)$ when the transition probability is 1 in polynomial time. Therefore, the problem of minimizing $AVG(A^-)$ is NP-hard when the transition probability is 1.

Theorem 2. *The influence minimization problem defined by (2) is NP-hard.*

Proof. Since the problem considered in Theorem 1 is a special case of minimizing $AVG(A^-)$ when the transition probability varies at edges, the general $AVG(A^-)$ minimization problem (2) is also NP-hard.

5.2 Approximation Algorithm

We design a simple approximation algorithm to find a set A of K nodes to solve (2). Initially $A = \emptyset$. We add K nodes into A to maximize $\sum_{v \in A} q \cdot \sigma(v)\text{over }_G$. The details of the algorithm are presented in Algorithm 1.

Algorithm 1. Approximation Algorithm(G)

Input:

- G: the contact network

Output:

- A: a set of K nodes for vaccination

```
function APPROXIMATION ALGORITHM($G$)
    Let $A = \emptyset$.
    Sort nodes in $V$ in a decreasing order of $q \cdot \sigma(v)\text{over }_G$.
    Add the first $K$ nodes in the sorted list into $A$.
    Output $A$;
end function
```

6 Simulation Study

6.1 Simulation Set-Up

We validate the proposed disease control strategy over a real-world data set from Facebook (http://snap.stanford.edu/data/egonets-Facebook.html), where facebook friend information of 3959 individuals are collected. In this simulation, we let

$$t_{ij} = \frac{|N_i \bigcap N_j| + 1}{|N_i \bigcup N_j| + 1}.$$

Here, N_i represents the set of nodes that are directly attached to i. The number '$1'$' appearing on the numerator and denominator is used to prevent p_{ij} from becoming 0. This action is based on the idea that a node always has a potential influence on any other node.

To verify the strength of the proposed strategy, we implement two other approaches for performance comparison. One approach employs a random alert strategy in the sense that a number of individuals are randomly chosen to be alerted for vaccination according to the number of available vaccines. This strategy has been widely used in the literature and is denoted as RD in this paper. Another implemented approach uses degree centrality, where the nodes with the largest node degrees are chosen for vaccination according to the number of available vaccines. It is argued by the previous work [18] that degree centrality yields promising results in predicting the risk of infection, compared to other centrality metrics.

We define the final infection ratio as the ratio of the total number of infected persons during the time of evaluation to the size of the entire population. The final infection ratio will be used as the primary performance metric for the evaluation of disease control strategies in our simulations.

We examine these strategies by varying different parameters such as the number of available vaccines, q, as well as the infection probability p mentioned in section 3.2. The initial infection ratio is defined as the total number of infected persons on the first day divided by the size of the population. The initial infected persons are chosen randomly in our simulations. We report our experimental results by an average of 50 runs.

6.2 Simulation Results

Fig. 1 reports the final infection ratio *vs.* the number of available vaccines, where we fix q to be 0.003, and the infection probability p to be 0.05. From the results, we observe that the final infection ratio declines as the number of available vaccines increases. When the number of people who receive vaccines is growing, more people are protected, and therefore the number of infected nodes is decreasing. Overall, the target vaccination strategies (degree and the proposed strategy) achieve a lower number of infected nodes than the random strategy, which is consistent with the previous work [15], [16], [3] whose strategies have resulted in less number of infections than the random strategy with the same cost of vaccines. The proposed strategy achieves the lowest final infection ratio when the number of available vaccines is less than 1000, and has almost the same final infection ratio as the degree strategy when the number of vaccines is larger than 1000. The results demonstrate that compared to the other two strategies, the proposed method protects more people with the same number of vaccines when there is a limited supply of vaccines.

In Fig. 2, we plot the final infection ratio *vs.* q, where the number of available vaccines is 200, and p is 0.05. Notice that for all three strategies, the final infection ratio increases as q increases. This is because when the number of vaccines is fixed, an increasing number of initial infected nodes will lead to a larger chance

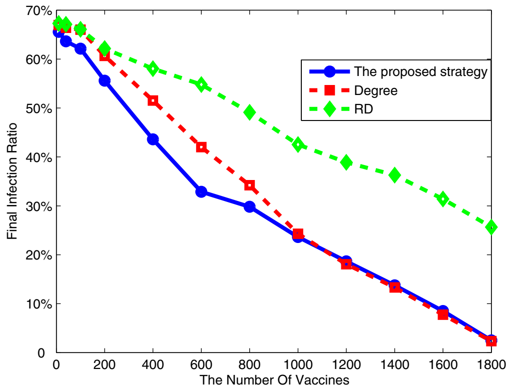


Fig. 1. Final Infection Ratio *vs.* Number Of Available Vaccines ($q = 0.003$, infection probability $= 0.05$).

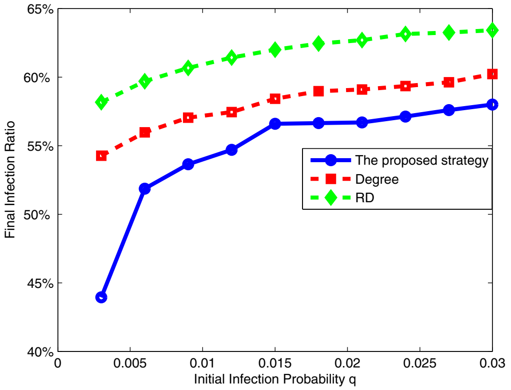


Fig. 2. Final Infection Ratio *vs.* q (Number of available vaccines $= 200$, infection probability $= 0.05$)

of spreading diseases to more nodes. In general, target vaccination strategies perform better than the random strategy, which is consistent with the results in Fig. 1. We also observe that the proposed strategy stands out from all others under different q. This indicates that our approach has a more effective capability to prevent disease from spreading no matter how seriously the disease starts initially in a population.

We also evaluate the vaccination strategies under different infection probabilities in Fig. 3, where the number of available vaccines is fixed to be 400, and q is fixed to be 0.003. It is observed that the final infection ratio increases when the infection probabilities grow larger. This is reasonable because for a single

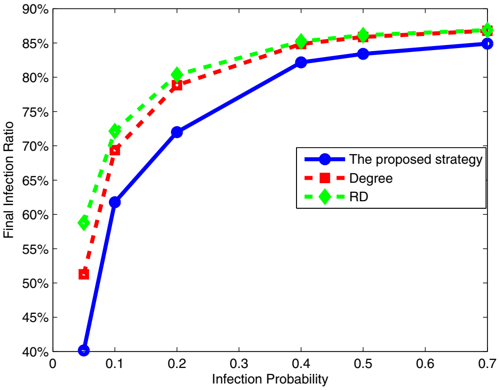


Fig. 3. Final Infection Ratio *vs.* Infection Probability (Number of available vaccines = 400, $q = 0.003$)

infected node, the number of nodes that might be infected by this node is a monotonically increasing function of the infection probability. Since $AVG(A^-)$ is the sum of the expected number of nodes infected by each node in A^-, it is also increasing as the infection probability ones. Our proposed strategy outperforms the other two under different p, again confirming the superiority of the proposed strategy to existing strategies. However, as the infection probability rises up, the superiority of the proposed strategy becomes less apparent. It can be observed that the performance differences among all the three strategies decline when the infection probability goes up. This reflects the real life scenario that when the disease is highly transmissible, there is less need to target people at risks, because everybody who has interactions with the infected person are all likely to be infected.

7 Conclusion

In this paper, we design a preventive disease control strategy to set up defenses against disease breakout ahead of time. The social communications are explored to extract social information such that we can determine the pattern of disease transmissions among individuals. We attempt to minimize the expected number of infected individuals by treating a small fraction of the population; this minimization problem is proved to be NP-hard in this paper and thus we propose an approximation algorithm. Simulations and comparisons are conducted to evaluate the performance of the proposed disease control strategy over a real-life data set. The results indicate that the proposed strategy is superior over existing ones.

References

1. Newman, M.E.J.: Spread of epidemic disease on networks. Phys. Rev. E 66, 016128 (2002)
2. Miller, J.C.: Spread of infectious diseases through clustered populations. Journal of the Royal Society Interface 6(41), 1121–1134 (2009)
3. Ren, Y., Yang, J., Chuah, M.C., Chen, Y.: Mobile phone enabled social community extraction for controlling of disease propagation in healthcare. In: 2011 IEEE 8th International Conference on Mobile Adhoc and Sensor Systems (MASS), pp. 646–651. IEEE (2011)
4. Dimitrov, N.B., Meyers, L.A.: Mathematical approaches to infectious disease prediction and control. In: Hasenbein, J.J. (ed.) Informs Tutorials in Operations Research, vol. 7, pp. 1–25 (2010)
5. Hethcote, H.: The mathematics of infectious diseases. SIAM Review 42(4), 599–653 (2000)
6. Anderson, R., May, R.: Infectious diseases of humans: dynamics and control. Oxford University Press (1991)
7. Zhou, T., Yan, G., Wang, B.H.: Maximal planar networks with large clustering coefficient and power-law degree distribution. Phys. Rev. E 71, 046141 (2005)
8. Meyers, L.A.: Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. Bull. Amer. Math. Soc. 44, 63–86 (2007)
9. Perisic, A., Bauch, C.T.: Social contact networks and disease eradicability under voluntary vaccination. PLoS Comput. Biol. 5(2), e1000280 (2009)
10. Huang, S.: Probabilistic model checking of disease spread and prevention. In: Scholarly Paper for the Degree of Masters in University of Maryland (2009)
11. Cha, M., Mislove, A., Gummadi, K.P.: A measurement-driven analysis of information propagation in the flickr social network. In: Proceedings of the 18th International Conference on World Wide Web, pp. 721–730. ACM (2009)
12. Goyal, A., Bonchi, F., Lakshmanan, L.V.: Learning influence probabilities in social networks. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 241–250. ACM (2010)
13. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146. ACM (2003)
14. Liang, X., Lu, R., Chen, L., Lin, X., Shen, X.: Pec: A privacy-preserving emergency call scheme for mobile healthcare social networks. Journal of Communications and Networks 13(2), 102–112 (2011)
15. Zhang, B., Cheng, X., Bie, R., Chen, D.: A community based vaccination strategy over mobile phone records. In: Proceedings of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare. mHealthSys 2012, pp. 2:1–2:6. ACM (2012)
16. Zhang, B., Gilani, S.M., Wu, D., Cheng, X., Bie, R.: Mobile phone based social relationship identification for target vaccination in mobile healthcare. In: Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones. PhoneSense 2012, pp. 5:1–5:5. ACM (2012)
17. Aspnes, J., Chang, K., Yampolskiy, A.: Inoculation strategies for victims of viruses and the sum-of-squares partition problem. Journal of Computer and System Sciences 72(6), 1077–1093 (2006)
18. Han, B., Hui, P., Kumar, V.A., Marathe, M.V., Shao, J., Srinivasan, A.: Mobile data offloading through opportunistic communications and social participation. IEEE Transactions on Mobile Computing 11(5), 821–834 (2012)

Author Index