A Distributed Clustering Algorithm for Voronoi Cell-based Large Scale Wireless Sensor Network

Jiehui Chen, Chul-soo Kim
Graduate School of Computer Science
Inje University
Gimhae, Gyungnam, South Korea
chenjiehui_0574@yahoo.co.jp,charles@inje.ac.kr

Fu Song
Institute of Software Engineering
East China Normal University
Shanghai, P.R.China
fsong@sei.ecnu.edu

***Abstract*—Due to resource constraints in Wireless Sensor Networks (WSNs), this paper contributes a distributed clustering algorithm suitable for a large scale Voronoi cell-based WSNs with sensors randomly deployed according to homogenous spatial Poisson process and each sensor becomes a cluster head (CH) with a possibility *p* while non-CH sensors join the cluster of the closest CH to form a Voronoi tessellation. We explore a new sensor node deployment and generate stochastic geometry for the proposed algorithm being capable of showing how the critical parameters give significant influences on minimizing energy cost. Without loss of generality, the highly creditable simulation results prove that the proposed algorithm outperformance the Max-Min D-Cluster algorithm in terms of energy efficiency under certain network specifications. Moreover, scalability and robustness of the algorithm are also verified over extensive experiments.**

Keywords- WSN, clustering algorithm, voronoi cell

I. INTRODUCTION

WSNs equipped with the extremely small, low cost sensors that possess sensing, signal processing and wireless communication capacities is highly capable of carrying out numerous tasks such as bio-chemical diffusion and military surveillance. Our objective is to create an efficient clustered WSNs with minimum energy cost. In literature, many clustering algorithms in various contexts have been proposed [1,2,3] aims at monitoring object boundaries by generating minimum number of clusters. However, many of them are heuristic and require time synchronization among the sensor nodes, which makes them only suitable for small WSNs. Moreover, to our knowledge, none of them is purposely for minimizing the energy cost in the network. [4] did minimize the total energy cost, but its assumption that each sensor node is aware of the whole network topology is theoretically impossible for large scale WSNs. In the Linked Cluster Algorithm (LCA)[5], a sensor node becomes a CH if it has the highest identity among all the one-hop sensor nodes or one-hop sensor nodes of its one-hop neighbors. The Max-Min d-Cluster Algorithm [6] generates *d*-hop clusters with a run-time of $O(d)$ round, and achieves better load balancing among the *CH*s, generates fewer clusters than [7]. Heinzalman *et al* [8] proposed a distributed algorithm for micro-WSNs where sensors elected themselves *CH*s with some probabilities and broadcast their decisions. But this algorithm only allows one-hop clusters to be formed, which might lead to a large number of clusters. And in their simulations, no evidence shows the optimal number of clusters in the proposed system. In the paper, we contribute a distributed clustering algorithm in the proposed multi-hop Voronoi cell-based WSNs.

The rest of this paper is organized as follows: explore a new sensor node deployment with creditable evidences and make general assumptions for the proposed algorithm followed by simple introduction to the network initialization phase in Section II. Then, from a mathematic view of point, derive stochastic geometry to form the algorithm for minimizing the energy cost in the network in section III. Section IV shows experiments conducted to monitor how the total energy cost changes to the changing values of critical parameters. Finally, conclude the paper in section □.

II. PRELIMINARIES

In this section, suppose that single sensors have no knowledge about the total number of sensors deployed and their corresponding locations. Instead, implement some mechanisms on the sink acting as a process center that might be somehow pregnable, easy to be compromised by adversaries. But the algorithm based on the premise that the safety of the sink is guaranteed at any price. At the first part of this section, we explore a new sensor node order for deployment proven to be better in terms of higher coverage and detectability. In the rest of this section, necessary assumptions are given for achieving the proposed algorithm and a brief introduction to the network initialization phase.

A. Explore a New Sensor Node Deployment

Here, give some basic definitions and notations throughput the paper. We model a multi-hop network by a undirected graph $G = (V, E)$ where V, $|V|=n$, is the set of wireless sensor nodes and there exists an edge $\{u,v\} \in E$, if and only if u and v can mutually receive each other's transmission. Namely, two sensor nodes are considered neighbours if the *Euclidean distance* is smaller or equal to the transmission rang r. the set of a sensor node $v \in V$ is denoted by €(v).

978-0-7695-3989-8/10 $26.00 © 2010 IEEE
DOI 10.1109/CMC.2010.230

ID: Every sensor node $v \in V$ in the network is assigned a unique identifier (ID) for identifying each other.

w_v: Every sensor node $v \in V$ in the network is assigned a weight w_v. In various applications of WSNs, sensor node weight plays an important role. Sometimes a high weight is required for the sake of redundancy or priority degree, while sometimes a high weight can be used to show the importance of its sending packets to the others. For the sake of simplicity, in the paper we stipulate that each sensor node has the same initial weight w_v=0.

Clustering is the whole procedure of partitioning the deployed sensor nodes into clusters, each cluster has a CH and its members.Each sensor node N_v becomes a CH with a probability p and broadcast its $\{ID_v, w_v\}$ as a *CH* to its $\epsilon(v)$ within its transmission range r and then the broadcasting message is forwarded to all the sensors at initial phase. Any sensor node, not itself a *CH* that receives such a broadcasting message joins the cluster of the closest CH.

Isolated sensor node: a sensor node that neither a *CH* nor has joined any cluster will be forced to become a *CH* after the clustering.

THEOREM 1. *Let $f\psi$ denote area coverage, namely the fraction of the geographical area that is in the sensing area of one or more sensors where sensor nodes can provide a valid sensing measurement and $\emptyset$ is the cartographic representation of area.Then,in Figure 1, get $\emptyset_{f(\beta)} \gg \emptyset_{f(\alpha)}$ in G = (V, E) where $E \neq \emptyset$.*

Proof: In literature, the majority of researches prefer *grid-based* (e.g. Figure 1(a)) geographic order for locating sensor nodes. Instinctively, get $\emptyset_{f(\beta)}$ is greater than $\emptyset_{f(\alpha)}$. Let's prove it with computational evidence as follow:

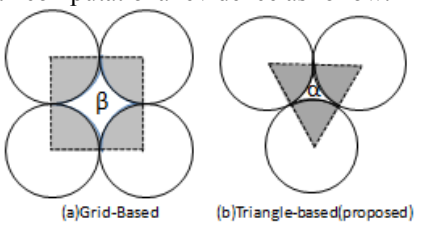

Figure 1. Explore a new approach of sensor node deployment based on area coverage

$$\emptyset_{f(\beta)} = (2r)^2 - 4\left(\frac{\pi r^2}{4}\right) = (4-\pi)r^2 \approx 0.86r^2 \quad (1)$$

$$\emptyset_{f(\alpha)} = \left(\sqrt{3} - \frac{\pi}{2}\right) r^2 \approx 0.1512r^2 \quad (2)$$

Since the calculation is very easy, allow me directly to the results. The difference is given by approximately $0.71r^2$. Even if the difference might be pretty small when r is small enough, for monitoring WSNs, accuracy is very sensitive. The smaller the value of $\emptyset_f$, the higher possibility that a moving object will not be detected.

THEOREM 2. *let d_v be a threshold distance ($d_v \gg r$) that is used for detecting sensor node N_v's $\epsilon(i)$. Get Triangle-based is more suitable for G = (V, E) where $E \neq \emptyset$,in terms of higher detectability.*

Proof: it's clear that the triangle-based has more detectable 1-hop $\epsilon(v)$ than grid-based at a rate 6:4 in quantity. Once detecting a task, N_v should relay detected *task messages* to another sensor node at a price of energy consumption. Denote H_v represents the total hops on the shortest routing path from N_v to the next candidate sensor node. Energy cost absolutely depends on H_v. Therefore, the problem shifted to prove that which one has more $\epsilon(v)$ within H_v distance for a consideration of detectability.

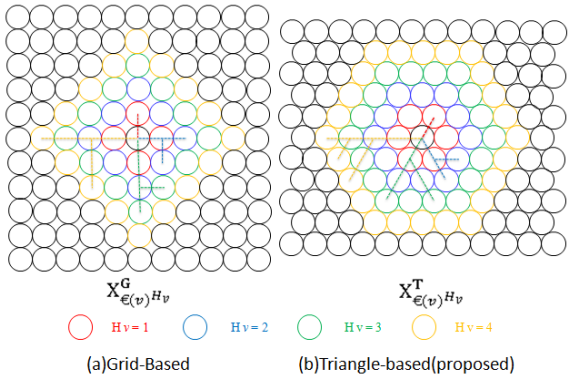

Fig. 2: Explore a new approach of sensor node deployment based on higher detectability

Let $X^T_{\epsilon(v)}$ and $X^G_{\epsilon(v)}$ denote the total number of detectable $\epsilon(v)$ at a H_v distance from N_v for Triangle-based and Grid-based sensor node deployment order respectively. Get:

$$X^T_{\epsilon(v)} = 3(1 + H_v)H_v \quad (3)$$

$$X^G_{\epsilon(v)} = 2(1 + H_v)H_v \quad (4)$$

The result is obvious that $X^T_{\epsilon(v)} \gg X^G_{\epsilon(v)}$ which inproves that Triangle-based is more suitable for G = (V, E) where $E \neq \emptyset$,in terms of higher detectability.

B. Assumptions

To achieve the proposed algorithm, give the following assumptions:

- Sensors with the same capabilities and functionalities. And the process center locates at the center of the network.
- Only the sensors on all the shortest routing paths forward the aggregated data.

- The communication environment is contention and error free, hence, no data retransmission needed. At the same time, ignore the time complexity of this algorithm.
- A distance between sensors and their CHs is measured by the number of communication hops H_v on the routing path, instead of geographic distance.

C. Network Initialization Phase

Step 1: After densely deployment according to the Poison process, all the normal sensors broadcast HELLO messages to its one-hop neighboring sensor nodes and store the information in its own *BN-array* [2] to make sure it's a boundary sensor node (BN) or not. In this way, all the boundary sensor nodes know their *boundary status.* There is a parameter in BN-array to show its *boundary status. Round*: assume that the process center (sink) is powerful enough to adjust its transmission power to produce sequential broadcasting HELLO messages to all the sensors deployed. Call the times of adjusting the transmission power as *Rounds.* e.g. the first round it's that the sink transmit it's one-hop reachable HELLO message to its one-hop neighboring sensor nodes; the second round should be the two-hop communication in a similar way; third…and so on. In Figure 2(b). there are 6 sensor nodes that received the broadcasting messages from the sink in the 1st *Round* need 2-hops to reach the sink at the center of the whole topology. In a similar way, 6×2 sensor nodes in 2nd *Round* need (2×2)-hops; (6×3) sensor nodes in 3rd *Round* need (2×3)-hops;…; $6\times R_i$ (R_i indicates the sequence order of *Rounds* needed for CapB algorithm to capture the boundary information.) sensor nodes in i^{th} *Round* need 2*i*-hops.

CapB Algorithm	
1.	Input G = (V, E) while E≠∅ do
2.	The sink tunes the power to achieve acceptable signal-to-noise radio and broadcast it to the sensors at i^{th} *Round* and be followed by $(i+1)^{th}$ *Round* transmission during a negligible time period.
3.	While receives two response messages from two adjacent *Round*s with the same time-slot duration in its buffer, the sink send a OK *confirm message* to the boundary sensor nodes, end Input
4.	Otherwise, back to 2

Parameters Setup	
n	The total number of sensors deployed
n_c	The number of sensors in a single cluster
D_{all}	The total length of segments, all the sensors->the sink
$D_{c\to s}$	The total length of segments, all level CHs->the sink
$\delta_{c\to s}$	The total energy cost, all level CHs->the sink
δ	Total energy cost of communicating gathered data from sensors to the sink through a hierarchy of CHs generated by the proposed algorithm.

Step 2: The sink starts communication within one-hop distance and then to two-hop and more. Here, non-boundary sensors that received the HELLO messages and already gave the response to the sink have no need to response again, while boundary sensors who knows its own boundary status in its own *BN-array* [2] have to response it every time until receiving a OK *confirm message* from the sink. The sink can calculate the time spent for sending a particular HELLO message and receiving a corresponding response message from a certain sensor, then record and calculate the data in the sink's buffer with a parameter of the *number of time slots* and location information of the sensors.Once, the sink receives two response messages from two adjacent *Round*s with the same time-slot duration in its buffer, the sink confirm that the response messages are from boundary sensor nodes. Then the sink stops sending HELLO messages and sending an OK *confirm message* at that transmission range, and then end.

III. TIPS ABOUT THE PROPOSED ALGORITHM

In this section, illustrate a single level energy-efficient clustering algorithm. Suppose that a single event is densely happened in a square area. We can capture the boundary information applying CapB algorithm. Tune the sink's power to achieve different transmission ranges. When the sink emits a radio within *r* distance transmission range, it get responses from all the sensors within 2*r*-distance after a unit time slot *t*. Repeat the same operation through varying the transmission power of the sink, finally, the sink receive two adjacent rounds' responses within the same time-slot duration *T*. Then, let's end the operation. Therefore, $T=R\times t$ (R indicates the total *Round* in *CapB* algorithm). Since network area is a nearly regular square area, the length of one boundary side should be equal to 4*rR*, calculate the $A=(4rR)^2$. Therefore, the number of sensors is a Poisson random variable with $E[n] = \lambda A$, Since the probability of becoming a *CH* is *p*, the *CH*s and non-*CH*s are distributed as per independent homogeneous spatial Poisson processes with intensity $\lambda_1 = p\,\lambda$ and $\lambda_0 = (1-p)\,\lambda$. To generate stochastic geometry for the proposed clustering algorithm and derive the optimal values of parameters for minimizing energy cost in the network without loss of generality.

By applying the above *CapB* algorithm, suppose a random sensor located at (x_i, y_i),*i*=1,2,…,n. Then get

$$E[D_{all}|N = n] = 12\sum_{i=1}^{R} i^2 = 2R(R+1)(2R+1) \qquad (5)$$

Since there are on an average *npCH*s with their locations independent, therefore, $D_{c\to s} = pD_{all} = 2R(R+1)(2R+1)p$. By arguments similar to [9], if N_v is a random variable denoting the number of PP0 process points in each Voronoi

cell (e.g. Figure 3) and L_v is the total length of segments connecting the PP0 process points to the nucleus in a Voronoi cell.

$$E\,[N_v|N=n]\approx E[N_v]=\frac{\lambda_0}{\lambda_1} \quad (6)$$

$$E\,[L_v|N=n]\approx E[L_v]=\frac{\lambda_0}{2\lambda_1^{3/2}} \quad (7)$$

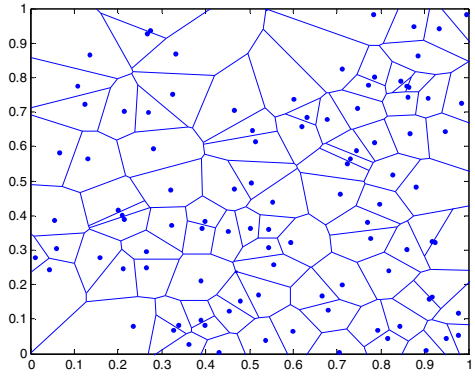

Figure. 3 Voronoi cell based WSN

Define $\boldsymbol{\delta_1}$ to be the total energy spent by all the sensors communicating 1 unit of data to their *CH*s, since there are on average $(2R)^2$ *CH*s, namely, $p(2R)^2$ Voronoi cells. Let assume that there exists very small amount of *isolated sensor node*s so that ignore them without any bad influence to the accuracy of the algorithm. Therefore, the expected value of $\boldsymbol{\delta}_1$ conditioned on N, is given by

$$E[\boldsymbol{\delta}_1|N=n]=np\,\frac{E\,[L_v|N=\alpha^2]}{r}=\frac{2(1-p)\,R^2}{r\sqrt{\lambda p}} \quad (8)$$

Conditioning on N, total energy spent by all the *CH*s communicating 1 unit of data to the sink is given by

$$E[\boldsymbol{\delta}_{c\to s}|N=n]=\frac{E[D_{c\to s}|N=\alpha^2]}{r}=\frac{2pR(R+1)(2R+1)}{r} \quad (9)$$

Then

$$E[\boldsymbol{\delta}|N=n]=E[\boldsymbol{\delta_1}|N=n]+E[\boldsymbol{\delta}_{c\to s}|N=n]$$

$$=\left[\frac{2(1-p)\boldsymbol{R}^2}{r\sqrt{\lambda p}}+\frac{2p\boldsymbol{R}(\boldsymbol{R}+1)(2\boldsymbol{R}+1)}{\boldsymbol{r}}\right] \quad (10)$$

$E[\boldsymbol{\delta}]$ is minimized by a value of p that is a solution of equation that gives partial derivative to (10) as follow:

$$\frac{2R(R+1)(2R+1)}{r}-\frac{R^2}{r\sqrt{\lambda p}}-\frac{R^2}{rp^{3/2}\sqrt{\lambda}}=0 \quad (11)$$

Then, get

$$-\mu p^{3/2}+p+1=0 \quad (12)$$

Where

$$\mu=\frac{2(R+1)(2R+1)}{R}\sqrt{\lambda}$$

The equation (12) has three roots, two of them are imaginary. The second derivative of the above function is positive only for the real root that is given by

Real Root:

$$\frac{1}{3\mu^2}-\frac{2^{\frac{1}{3}}(-1-6\mu^2)}{3\mu^2(2+18\mu^2+27\mu^4+3\sqrt{3}\,\mu^3\sqrt{27\mu^2+4})^{\frac{1}{3}}}+\frac{(2+18\mu^2+27\mu^4+3\sqrt{3}\,\mu^3\sqrt{27\mu^2+4})^{\frac{1}{3}}}{2^{\frac{1}{3}}(3\mu^2)} \quad (13)$$

Hence, if and only if the value of p is equal to the real root, the algorithm does really minimize the energy cost. So far, we derived equations for computation of optimal values of dependable parameters to measure the proposed algorithm. Without numerical simulation results, we cannot prove it's accuracy and robustness. In the next section, evidences will be given in the form of numerous simulation results.

IV. Simulation and Numerical Results

In this section, we simulated the proposed algorithm with totally n distributed sensors in a square of 1000 sq. units using VC++ programming. Energy dissipation follows Low Energy Adaptive Clustering Hierarchy (LEACH) protocol. The experiments were conducted with the communication range r was assigned to be 1 unit and total number of sensors n is assigned to be 400, 1600, 2500 with R=10, 20, 25 respectively. Moreover, the processing center is assumed to be at the center of the network area. Don't consider the unexpected errors and influences from outside circumstance.

For the simulation experiments, considered a range of possible value of the probability (p) less than 0.1 for most of potentials. For each of possible value of p, compute the density of Poisson process λ for generating the network under different network conditions. The results are provided in Figure 4. In figure 4, CapB algorithm was used to detect the boundary of the network with R=10, R=20 and R=25 respectively. Then vary the value of the density of Poisson process (λ) to get the willing values of p for computation on minimized energy cost ($\boldsymbol{\delta}$). However, it shows that the value of p decreases as the value of λ increases stably at an interval {0.03, 0.1}. To achieve p with a value of smaller than 0.03, we have to manage the rapidity of changing λ at a high value since clustering algorithm are well working in a densely deployed large scale WSNs, while to achieve p in excess of 0.1, we don't need to concern too much because there are few sensors randomly distributed in such a large scale area with λ pretty small that indicates sensors are difficult to get communicate with each other, they are of great potential to be geographically separated. In this case,

the algorithm produce huge amount of *isolated sensors* that is object to the assumption and beyond our consideration.

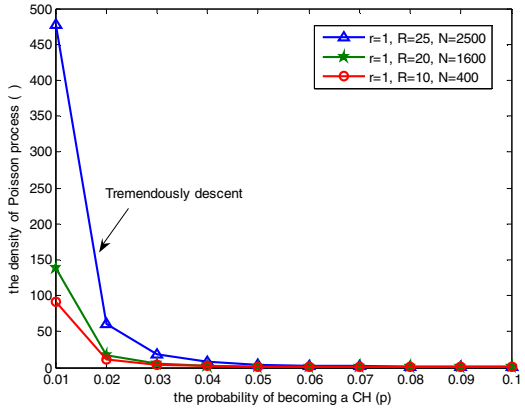

Figure 4. The computation of parameters {p, }

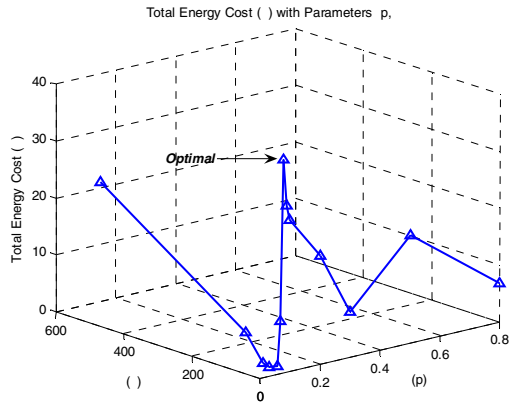

Figure 5. Optimal value p for minimizing total energy cost (δ)

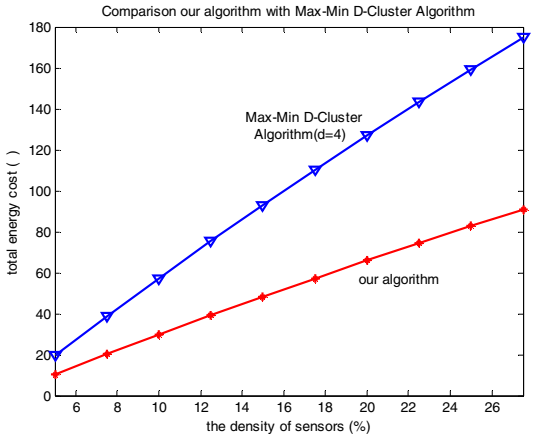

Figure 6. Comparison with Max-Min D-Cluster algorithm

Each data point in Figure 5 corresponds to the average energy cost over 100 experiments. It is verified that the energy spent in the network is indeed minimized at the theoretically optimal value of p at "0.08" under a network condition of {r=1, R=10, N=400} in a randomly distributed large scale Voronoi cell based WSNs. The optimal value of p here will be of more considerable for the future research. Now, let's do comparative study between popular Max-Min D-Cluster algorithm and the proposed clustering algorithm in terms of minimizing energy cost.

In Figure 6. the pre-obtained optimal values of all the critical parameters of the proposed algorithm in simulation model are used to evaluate the performance of the algorithm. At same time, we evaluated the Max-Min D-Cluster Algorithm with d=4 (already proven to be efficient). The result (e.g. Figure 6) clearly verifies that the algorithm performances better in terms of energy cost in the network under this network specifications.

V. CONCLUSION

In the paper, a distributed clustering algorithm was proposed for organizing sensors in a large scale Voronoi cell based WSNs with an objective of minimizing the total energy cost. The optimal values of the critical parameters of our algorithm were found in forms of math equations over numerous times simulations. However, we are facing problems to make all the assumptions available since the algorithm really has a time complexity O(k_n) in a contention-free network that are critical for a large scale WSN. In near future, we intend to explore a hierarchical clustering algorithm that might be more efficient and capable for more complex monitoring WSNs.

REFERENCES

[1] C.R.Lin and M.Gerla,"Adaptive Clustering for Mobile Wireless Networks", *Journal on Selected Areas in Communication*, Vol. 15 pp. 1265-1275, September 1997.

[2] Jiehui Chen and Mitsuji Matsumoto, "EUCOW: Energy Efficient Boundary Monitoring for Unsmoothed Continuous Objects in WSN", *The Sixth IEEE International Conference on Mobile Ad-hoc and Sensor Systems, IEEE WAASN09 in conjunction with IEEE MASS09*, Macau, China , Oct., 2009.

[3] W.R.Heinzelman, A.C.and H.Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks", in Proceedings of *IEEE HICSS*, January 2000.

[4] C.F.Chiasserini, I. Chlamtac, P. Monti and A. Nucci, "Energy Efficient design of Wireless Ad Hoc Networks", in Proceedings of *European Wireless*, February 2002.

[5] D.J.Baker and A.Ephremides,"The Architectural Organization of a Mobile Radio Network via a Distributed Algorithm", *IEEE Transactions on Communications*, Vol. 29, No. 11, pp. 1694-1701, November 1981

[6] A.D.Amis, R.Prakash, T.H.P.Vuong and D. T. Huynh, " Max-Min D-Cluster Formation in Wireless Ad Hoc Networks", in Proceedings of *IEEE INFOCOM*, March 2000.

[7] A. Ephremides, J.E. W. and D.J.B., "A Design concept for Reliable Mobile Radio Networks with Frequency Hopping Signaling", Proceeding of *IEEE*, Vol. 75, pp. 56-73, 1987.

[8] W.R.Heinzelman,A.C.and H. Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks", in Proceedings of *IEEE HICSS*, Jan. 2000.

[9] S.G.Foss and S.A.Zuyev, "On a Voronoi Aggregative Process Related to a Bivariate Poisson Process", *Advances in Applied Probability*, Vol. 28, no. 4, pp. 965-981,1996.