IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

Energy Efficient Clustering Protocol for Heterogeneous Wireless Sensor Network: A Hybrid Approach using GA and K-means

Shashi Bhushan
School of Computer and Information Science, IGNOU
New Delhi, India
s.bhushan2k5@gmail.com

Raju Pal
Dept. of CSE
Jaypee Institute of Information Technology, Noida, U.P., India
raju3131.pal@gmail.com

Svetlana G Antoshchuk
Institute of Computer System
ONPU
Odessa, Ukraine
asgonpu@gmail.com

***Abstract*—A hybrid approach combining genetic algorithm(GA) and K-means algorithm, called KGA is proposed in this paper for design of clustering protocol with energy efficiency for non-homogeneous wireless sensor network. The problem of optimal clustering can be considered as a problem for searching for an optimal number of clusters in a big search space such that WSN metrics are optimized. In the proposed protocol, distance between clusters, distance within clusters and a number of cluster heads are employed to search for optimal number of clusters and cluster heads. Maximization of energy saving and lifetime of a network are the two important metrics. The KGA is designed with a hybrid approach to population initialization scheme and objective function. The superiority of the protocol over other heuristic and meta-heuristic techniques is extensively demonstrated on several parameters: energy efficiency, network life time and throughput**

Keywords— Optimal clustering, WSN, Genetic Algorithm, K-means.

I. INTRODUCTION

A WSN is a large network of sensor nodes characterized by sensing capability, limited wireless range for communication, limited processing power, limited storage and battery capacities. Sensor nodes are deployed in a large Sensor nodes sense physical properties of environment such as humidity, temperature, sound, etc., collect the sense data and forward it to the base station for further processing. They have potential application in large domains: environment monitoring, disaster warning system, health system and military application [1].This has been enabled by the availability of cheaper and smarter sensors. One of the major challenge in WSN, however, is how to maximize the energy saving and lifetime of the network.

Clustering is one of the most efficient technique for maximizing energy saving and increase the lifetime of the network. K-means is one of the most popular data clustering techniques which partitions the data into K-clusters such that there is more similarity within clusters but less similarity between clusters [2]. Within each cluster there is a node called a cluster head which aggregates data received from sensor nodes within a cluster and forwards it to the base station. Clustering has been the most popular approaches for scalability of the network as well as designing energy efficient routing protocol for WSN. Clustering and routing with energy efficiency are the two widely used problems in WSN which belong to optimization class of problems. A number of heuristic and mea-heuristic techniques have been proposed for the above problems with the objectives of extending the lifetime of a network and maximize energy saving. LEACH [3] and SEP [4] are the two main prominent heuristic techniques for clustering routing in WSN. All sensor nodes in LEACH are homogeneous whereas in SEP, some nodes maintain heterogeneity in terms of energy. LEACH utilizes randomized rotation of CHs for proper load distribution of energy loads among all the nodes in a cluster. However it depends upon the probability model. In SEP, CH selection probabilities of nodes are weighted by initial energy of each node compared to the others. In[5] a meta-heuristic technique using GA[6] is used to generate energy efficient hierarchical clusters. Its fitness function include several distance parameters that is ,distances between sensor nodes and the base station, sensor nodes and a cluster head and cluster heads and a base station as well as standard deviation in cluster distance, transfer energy and number of transmissions. In [7], three cluster quality parameters have been added to derive fitness function in order to measure quality of energy efficient clustering which includes: intra-cluster distance, inter- cluster distance and a number of CHs. Higher CHs result in more energy consumption compared to few CHs. In [8],GA based protocol is proposed to extend the lifetime of a network and improve the stability period by finding optimal number of CHs and their locations based on minimizing the energy consumption of all the sensor nodes. PSO [10] is applied for both clustering and routing algorithms in [9]. For finding out a routing path, there is a tradeoff between transmission distance and a number of hop counts for selection of a routing path. For the clustering algorithms, all the CHs which are in use heavily as a next hop relay node in data forwarding are assigned a few nodes to balance the energy consumption. In [11] GA is used for both clustering and routing algorithms. Residual energy of the gateways and distance from sensor nodes to their CHs are used for clustering whereas for selecting the routing path, the residual energy of the gateways along with a tradeoff between transmission distance and a number of hops are used. Unlike the above protocols where population initialization is random, the proposed protocol follows the deterministic approach to population initialization. The population is seeded with K-means algorithm to have good quality genes in the beginning. GA is further used to form clustering.

The main contribution of the papers is the following:

• Population initialization and clustering through a mix of deterministic and random approaches.

• Performance comparison with high performance heuristic and meta-heuristic techniques

The rest of the paper is organized as follows: Section 2 describes GA, section 3 formulates a problem, section 4 gives a detailed description of the proposed algorithm, KGA, and Section 5 presents analysis of the result. Section 6 concludes the paper with future work.

II. PRELIMINARIES

A. An Overview of Genetic Algorithm (GA)

GA is an optimization technique based on the principles of evolution and natural genetics and randomized search. It performs search operation in very complex, large and multimodal landscape in order to provide a near optimal solution to a problem. Initially population representing individual solutions is generated randomly in a search space. Each solution is a chromosome. A fitness function is used to measure quality of solutions in the population. Through use of biologically inspired operations like crossover and mutation on the selected population, a new generation of a good solution is selected. This process (of selection, crossover and mutation) is continued for fixed number of iterations or till a termination condition is fulfilled. The following algorithm explains steps in writing GA.

Algorithm 1: Genetic Algorithm
```
t = 0               ;first iteration
initialize P(t)     ;initial population
evaluate P(t) ;evaluate population using fitness function
while ( termination_ condition not fulfilled ) do
    t = t+1
    choose P(t) from P(t-1); based on the evaluation result
    modify P(t);using cross-over and mutation operators
    evaluate P(t)
End while
```

B. Energy Model

The following equations are used to compute the transmission and reception energy consumption as proposed in [7]

$$E_{Trans}(k, dist) = E_{elec} \cdot k^{\varepsilon_{fs}} \cdot k \cdot d^2, dist \le d_0 \quad (1a)$$

$$E_{Trans}(k, dist) = E_{elec} \cdot k + \varepsilon_{mp} \cdot k \cdot d^4, dist > d_0 \quad (1b)$$

$$E_{Rcv}(k, dist) = E_{elec} \cdot k \quad (2)$$

where E_{Trans} (l, d) is k bit transmission energy consumption at a distance of dist , E_{Rcv} (k, dist) is receiving energy consumption for k bits data. Several factors such as encoding and modulation techniques and filtering influence E_{elec} which is the consumed energy by circuits while it is transmitting and receiving data. ε_{fs} and ε_{mp} refer to energy consumption required by the amplifier for the free space and multipath fading respectively which depend upon (i) distance between a transmitter and a receiver and (ii) an acceptable bit error rate.

III. KGA

This section covers the details of proposed method namely KGA from the initials of K-means and GA. It is a new hybrid evolutionary algorithm combining GA and K-means to solve clustering problem of WSN. Most of the routing protocols based on GA choose random population but these may not be efficient for sparsely deployed sensor networks because it may lead to the selection of CHs with low density. This may affect the final clustering result. A solution to this limitation is selection of good quality CHs in the population. A combination of meta-heuristic approach for supporting high exploration and deterministic approach for selection of good quality CHs can provide better clustering solution for HWSNs. In the proposed method, the population of GA is seeded with K-means in order to get good clustering result.

A. Problem Statement

The problem being studied in this paper is how to achieve optimal number of clusters in a large heterogeneous sensor network in order to maximize its lifetime by minimizing the energy consumption . Optimal clustering is an NP hard problem [12]. Various meta-heuristic protocols are available in literature to solve such kind of problems. Genetic algorithm is the most popular meta-heuristic method. In this paper a novel clustering method based on GA is proposed which considers the clustering parameters: intra-cluster distance, inter-cluster distance and number of clusters.

The main objective of clustering WSNs is to minimize the overall energy consumption in the network. Energy consumption of a node is directly proportional to the communication distance between the nodes. Long distance transmission between sensor nodes and a base station always consumes more energy. Energy consumption required in transmitting data by a sensor node S_i to the corresponding CH_i is represented as:

$$E_{Trans}(s_i, CH_i) \alpha d^{\lambda}. \quad (3)$$

Where d is the distance between sender and receiver and λ is a path loss component, whose value lies between 2 and 4,ie $2 \le \lambda \le 4$. If d is minimized, the total intra-cluster distance is also minimized. Then, E_{Trans} is also minimized, resulting in increasing the network lifetime.

Efficient clustering plays a key role to minimize the communication distance between the nodes. To improve the quality of clustering a novel fitness function is proposed which is described in section III-C. Moreover, to overcome the problem of long distance communication between CH to sink, multihop routing is also considered. Multihop communication is formulated as follows:

i. WSN = { C_1, C_2 … C_N } ; set of clusters

ii. C_i = { S_1, S_2 …. S_j }; set of sensor nodes. *j* is $|C_i|$

iii. dist (C_i , C_j) ; distance between two CHs

iv. next-hop(C_i) = { C_j | C_j ∈ dist (C_j ,BS) < dist (C_i , BS)

B.Chromosomes Structure and Population Initialization

The performance of GA is based on the efficient population initialization. In KGA, K-means is used to seed the population so that better CHs can be considered even in the case of sparsely deployed sensor network. The cluster-heads returns by K-means are represented as 1s in a chromosome and the rest of the sensor nodes are represented by 0s denoting normal sensor nodes. Dead nodes are represented as -1s. The detailed algorithm is given in Algorithm 2.

Algorithm 2 Population Initialization
For each chromosome i in Population P
  For each node j in a chromosome i
    If node(j).energy>0
      k=round((Alive nodes*0.2);
    Else if k>0
      ids=apply K-means with sensor nodes and k
      chromosome(i,j)=1 if i∈ ids
      chromosome(i,j)=0 if otherwise
    End If
End for
End for

B. Fitness Function

In KGA, the fitness function is defined based on intra and inter cluster distances and the number of alive nodes in the network. The objective is to minimize it. The function is defined in Eq. (7).

$$\text{Minimize } f = \alpha(D_{intra}/D_{inter}) + \beta * |WSN|, \qquad (4)$$

where, α and β are weighted coefficients such that α+β=1. D_{intra} and D_{inter} are the intra-cluster and inter-cluster distances given in Eq. (5) and (6) respectively and $|WSN|$ gives the number of cluster heads in the network given in Eq. (4).

$$D_{intra} = \sum_{i=1}^{CHs} \sum_{\forall S_j \in C_i} d(S_j, CH_i), \text{ where } j = 1,...,n . \qquad (5)$$

$$D_{inter} = \sum_{\forall C_i, C_j, C_i \neq C_j} d(CH_i, CH_j). \qquad (6)$$

C. Clustering Algorithm

GA based clustering algorithm in KGA has two main phases: (i) Setup phase which is a clustering formation phase and (ii) Steady state phase in which intra-cluster and inter-cluster communication takes place through multi-hop communication. In set up phase cluster head selection is guided by GA. The steady state is of our protocol is similar to LEACH. The details steps are defined in Algorithm 3.

IV. Result Analysis

In this section the performance of the proposed protocol is analysed against other existing routing protocols as SEP, IHCR and ERP in terms of network lifetime, residual energy and throughput with 10% and 20% advanced nodes. The simulation is performed in MATLAB 2016a. The discussion on each considered performance metrics is given in following subsections.

Algorithm 3 KGA
Setup Phase
1. Initialize Parameters of WSN fields
2. Initialize Heterogeneous WSN and Energy Model
3. Creation of Random Sensor Network
4. Initialize Evolutionary Algorithm Parameters
5. Seed the population of GA with K-means
6. Perform GA based clustering
7. Select the best CHs based on fitness value
8. CHs broadcast a message to all the remaining sensor nodes
9. The nodes select their CHs based on distance
10. Formation of CHs is complete in the first round

Steady State Phase
Sensor nodes start sensing the environment and transmitting data to their CHs as per their TDMA schedule. After receiving data, CHs aggregate and send data to the BS in multi-hop manner.

A. Network Lifetime

A Network lifetime can be shown by capturing the number of alive nodes at each round till every node in the network dies. Figure 1 and Figure 2 depict the comparative network life of each considered algorithm over 10% and 20% of node heterogeneity respectively. In each of the scenario KGA outperforms the other considered protocols.

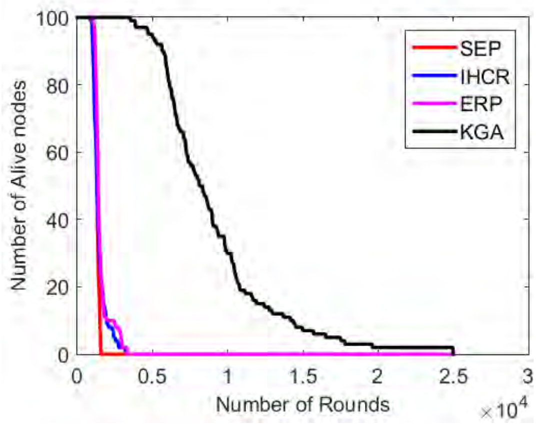

Fig. 1. Network Lifetime (10% advanced node)

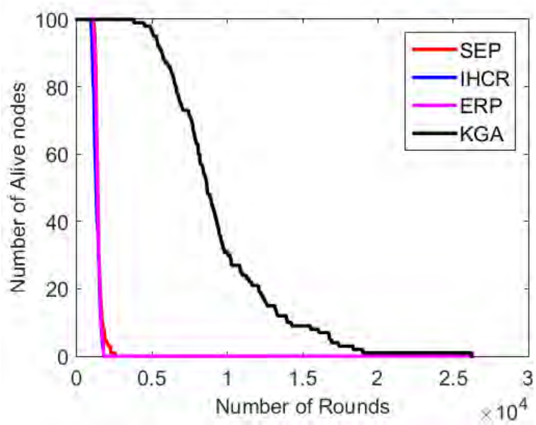

Fig. 2. Network Lifetime (20% advanced node)

To clearly depict the effectiveness of proposed algorithm, quantitative analysis has been performed and presented in Table 1 and 2 for 10% and 20% of advanced nodes(AV) in the network respectively. It can be observed from the table 1 that in KGA 10 nodes died in 5623 rounds while in other protocols, nodes died very early as 1153, 1047, and 1205 round for SEP, HCR, and ERP respectively. Similar in Table 2, nodes stay longer in KGA as compared to other protocols. Further, Table 3 and Table 4 have been presented to visualize the depletion of normal nodes(NRs) in comparison with advanced nodes(AVs) in the network. First 6000 round have been considered for comparative analysis and network statistics have been captured in different intervals of rounds like 10%, 20% and so on. Table 3 represents the statistics for the network with 10% advanced nodes. From the table, it is visible that almost all nodes died in HCR and ERP till 3000 rounds while in KGA no advanced node has died till 6000 round. Similarly, Table 4 represents the statistics for the network with 20% advanced nodes. From the table, it can be seen that in KGA, no advanced nodes have died in 6000 rounds.

TABLE I. NETWORK LIFETIME WITH 10% ADVANCED NODES

M=0.1				
% nodes died	SEP	HCR	ERP	KGA
10	1153	1047	1205	5623
20	1204	1099	1256	6130
30	1232	1159	1295	6602
40	1278	1255	1364	7263
50	1300	1303	1390	8097
60	1328	1372	1432	8972
70	1370	1459	1535	9943
80	1445	1940	1683	10788
90	1494	1956	2445	14170
100	1563	3220	3317	25002

TABLE II. NETWORK LIFETIME WITH 20% AVS

%dead nodes	SEP	HCR	ERP	KGA
10	1185	1050	1190	5648
20	1219	1146	1258	6540
30	1250	1208	1312	7605
40	1284	1276	1364	8071
50	1323	1353	1408	8648
60	1368	1430	1480	9288
70	1424	1569	1572	10034
80	1529	1928	1887	12085
90	1791	2529	2747	13971
100	2236	3536	3673	262389

TABLE III. ROUND HISTORY OF DEAD NODES WITH 10% AV AND NR NODES FOR A TOTAL OF 6000 ROUNDS

%rounds	HCR		ERP		KGA	
	AV	NR	AV	NR	AV	NR
10	0	0	0	0	0	0
20	0	35	0	9	0	0
30	0	86	0	89	0	0
40	4	90	0	90	0	0
50	8	90	7	90	0	0
60	_	_	_	_	0	1
70	_	_	_	_	0	3
80	_	_	_	_	0	5
90	_	_	_	_	0	8
100	_	_	_	_	0	16

TABLE IV. NUMBER OF DEAD NODES WITH 20% AV AND NR NODES FOR A TOTAL OF 6000 ROUNDS

%rounds	HCR		ERP		KGA	
	AV	NR	AV	NR	AV	NR
10	0	0	0	0	0	0
20	0	20	0	4	0	0
30	0	80	0	78	0	0
40	2	85	0	80	0	0
50	6	85	5	80	0	0
60	_	_	_	_	0	0
70	_	_	_	_	0	1
80	_	_	_	_	0	2
90	_	_	_	_	0	8
100	_	_	_	_	0	13

B. Residual Energy

Figures 3 and 4 shows the comparison of KGA with the other protocols in terms of RE versus number of rounds with 10% and 20% advanced nodes respectively. There is a less steepness of the curve due to fairness in the energy load distribution and gradual dissipation of energy in the proposed protocol. The result is further validated through Tables 5-6.

C. Throughput

Figures 5 and 6 demonstrate the number of data packets sent to the base station by CH nodes per round with 10% and 20% node heterogeneity respectively. From these figures it can be observed that in KGA cluster heads send more packets to the base station compared to the other protocols.

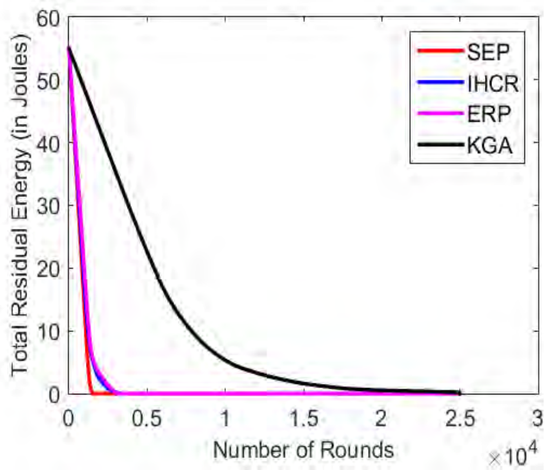

Fig. 3. Residual energy with 10% advanced nodes

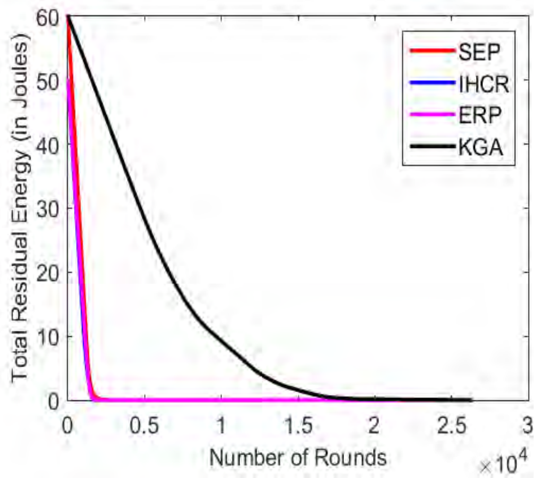

Fig. 4. Residual energy with 20% advanced nodes

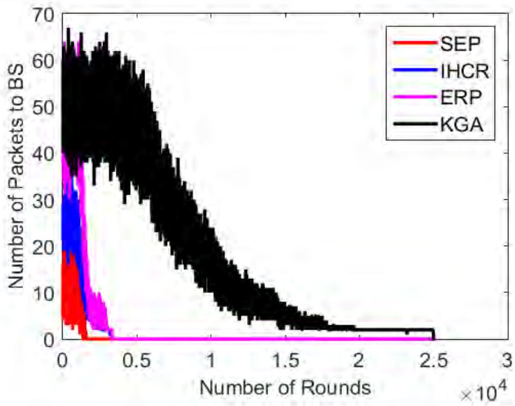

Fig. 5. Throughput (10% advanced nodes)

TABLE V. RESIDUAL ENERGY WITH 10% ADVANCED NODES

%rounds	SEP	HCR	ERP	KGA
10	27.66	26.89782	28.28127	51.05516
20	4.33	5.493893	6.849335	47.11319
30	_	2.014	3.06148	43.15388
40	_	0.088	0.986013	39.19862
50	_	0.002	0.229928	35.25073
60	_	_	_	31.31175
70	_	_	_	27.49672
80	_	_	_	23.78303
90	_	_	_	20.23968
100	_	_	_	16.91759

TABLE VI. RESIDUAL ENERGY WITH 20% ADVANCED NODES

%rounds	SEP	HCR	ERP	KGA
10	29.72	31.86016	33.49057	56.16828
20	5.11	10.43817	12.21157	52.34371
30	_	3.06148	3.82058	48.51257
40	_	0.986013	1.654297	44.68288
50	_	0.229928	0.158129	40.85083
60	_	_	_	37.00909
70	_	_	_	33.22163
80	_	_	_	29.49268
90	_	_	_	25.95671
100	_	_	_	22.78291

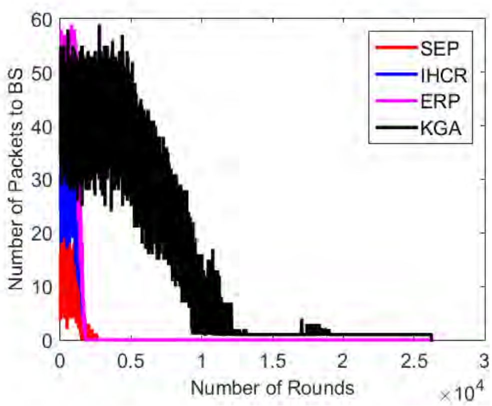

Fig. 6. Throughput (20% advanced nodes)

V. CONCLUSION

Clustering and routing are two important aspects in WSN. GA has been extensively used to solve problems related to these aspects. The novel idea presented in the proposed work is related to hybridization of population initialization. In this paper a hybrid approach combining GA and K-means is proposed to do clustering of WSN. Initial population of GA is seeded with K-means to have good quality CHs .The fitness function is defined over the parameters like intra-cluster distance, inter-cluster distance and number of clusters .The experimental results show that the performance of the proposed protocol is better than IHCR, SEP and ERP in terms

of network life time, residual energy with 10% and 20% advanced nodes and throughput. In the future, the proposed work will be tested on multiple WSNs and a large population size.

REFERENCES

[1] Ian F. Akyildiz, et al., "A survey on sensor networks," IEEE Communications magazine, vol. 40.8, pp. 102-114, 2002.

[2] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, 2 ed. Morgan Kaufman Publishers, 2006

[3] W. B. Heinzelman, A. P. Chandrakasan, H. Balakrishnan, “An application-specific protocol architecture for wireless microsensor networks,” IEEE Transactions on wireless communications, vol. 1(4) pp. 660-670, Oct. 2002

[4] Georgios Smaragdakis, Ibrahim Matta, and Azer Bestavros, “SEP: A stable election protocol for clustered heterogeneous wireless sensor networks,” Boston University Computer Science Department, May 31 2004.

[5] A. W. Matin, and S. Hussain, “Intelligent hierarchical cluster-based routing,” in: Proceedings of the international workshop on mobility and scalability in wireless sensor networks (MSWSN) in IEEE international conference on Distributed Computing in Sensor Networks (DCOSS), pp. 165–172, 2006.

[6] Zbigniew Michalewicz, Genetic algorithms + data structures = evolution programs. Springer, 2009

[7] Bara’a A. Attea, and Enan A. Khalil. “A new evolutionary based routing protocol for clustered heterogeneous wireless sensor networks,” Applied Soft Computing vol. 12.7, pp. 1950-1957, 2012.

[8] Mohammed Abo-Zahhad, et al. "A new energy-efficient adaptive clustering protocol based on genetic algorithm for improving the lifetime and the stable period of wireless sensor networks," International Journal of Energy, Information and Communications, vol. 5.3 pp. 47-72, 2014.

[9] Pratyay Kuila, and Prasanta K. Jana, “Energy efficient clustering and routing algorithms for wireless sensor networks: Particle swarm optimization approach,” Engineering Applications of Artificial Intelligence, vol. 33 pp. 127-140, 2014.

[10] James Kennedy, "Particle swarm optimization." Encyclopedia of machine learning. Springer US, 2011. 760-766

[11] Suneet K. Gupta, and Prasanta K. Jana. “Energy efficient clustering and routing algorithms for wireless sensor networks: GA based approach,” Wireless Personal Communications, vol. 83.3 pp. 2403-2423, 2015.

[12] Stefano Basagni, et al., “A generalized clustering algorithm for peer-to-peer networks,” in Workshop on Algorithmic Aspects of Communication. 1997.