Optimal cluster sizes for wireless sensor networks: An experimental analysis

Anna Förster[1], Alexander Förster[2] and Amy L. Murphy[3]

[1] Universitá della Svizzera Italiana, Switzerland
[2] IDSIA, Manno, Switzerland
[3] FBK-IRST, TN, Italy

Abstract. Node clustering and data aggregation are popular techniques to reduce energy consumption in large WSNs and a large body of literature has emerged describing various clustering protocols. Unfortunately, for practitioners wishing to exploit clustering in deployments, there is little help when trying to identify a protocol that meets their needs. This paper takes a step back from specific protocols to consider the fundamental question: *what is the optimal cluster size* in terms of the resulting communication generated to collect data. Our experimental analysis considers a wide range of parameters that characterize the WSN, and shows that in the most common cases, clusters in which all nodes can communicate in one hop to the cluster head are optimal.

1 Introduction

Clustering in wireless sensor networks (WSNs) is the process of dividing the nodes of the WSN into groups, where each group agrees on a central node, called the cluster head, which is responsible for gathering the sensory data of all group members, aggregating it and sending it to the base station(s). One of the first, and most well-known clustering approaches is LEACH [1], which relies on randomly selected cluster heads and network-wide broadcasts to assign nodes to cluster heads. Since this initial step, many clustering protocols have been proposed distinguishable according to whether the resulting cluster are: random or location based; one or multiple hops wide; randomly or intentionally shaped, e.g., as squares, etc.

Despite the wealth and variety of protocols and their individual evaluations through simulation, experimentation, and comparison to existing approaches, little help is available for the WSN application developer who must answer the question "*what is the optimal cluster?*" We define the *optimal* cluster as the one sized such that routing data from the cluster members to cluster heads and subsequently to base stations incurs the minimal communication overhead.

We first confronted this question of optimal clusters during our development of Clique [2], a low energy, low overhead clustering protocol that identifies tunable-size clusters based on geographic information. While evaluating the communication overhead of various cluster sizes, we observed that the optimal cluster size for a given network is complex, depending on a wide range of parameters.

At the time, as protocol developers, we identified optimal clusters through hundreds of simulated experiments. *Practitioners*, instead, require a more general approach that, during the deployment phase, guides them in the identification of applicable clustering algorithms in parallel with the fine-tuning of the network and clustering parameters to yield minimal communication overhead.

Recently the presentation and evaluation of new communication protocols and algorithms for WSNs has faced criticism [3–5]. Regarding clustering for WSNs, extensive research time has been invested to develop energy-efficient, low overhead protocols, relying on experimental and/or theoretical analysis and comparison to pre-existing protocols. While this approach is valid, its implementation often suffers, resulting in incomplete evaluations. Many different assumptions are made about the network properties such as size, number of nodes, communication reliability, etc. The experiments are performed with different simulators or testbeds and the results are hardly comparable across articles. This evaluation practice results in a jungle of protocols, separated in two groups: those developed for real WSNs that have been shown to be practical, and those that have been evaluated only in simulation. Important for the first group is that *it works*. Driven by the requirement of simple implementation, practitioners typically select and implement only one or two protocols of the same type. However, in their choice of protocols they rely on the theorists, specifically their developed algorithms and their analyses. The main problem with these theoretically developed protocols is that comparison to one another is extremely difficult as they use very different, poorly documented simulation setups. Thus, practitioners select a protocol based on unreliable information or very often decide to develop their own protocol to ensure it meets their requirements. Unfortunately, however, having a protocol that "works" does not automatically imply that it is the best fit.

Defining and comparing to such best-fit, *optimal* solutions is common when developing new protocols. For example, routing is often compared against the shortest path, which can be easily computed, and thus offers a clear comparison point for new protocols. Instead, for *optimal clusters*, there is neither a clear definition nor a standard algorithm for its computation as the number of parameters affecting the result is large. Therefore, in this paper, we explore the entire problem of optimal clustering in detail, through both the lenses of *analysis* and *experimentation*. First, Section 2 sketches the current state of the art of clustering algorithms and their evaluations, motivating the need for further analysis. Our theoretical analysis begins in Section 3 by defining the parameter space and the optimality of clusters in terms of communication costs. Section 4 then explores the parameter and property space of clustering (*analysis*) and identifies solutions (*experimentation*). Section 5 summarizes our findings and offers directions for future research.

TAXONOMY	Examples	# HOPS IN CLUSTER	CLUSTER HEADS	CLUSTER FORM
RANDOM	LEACH [1], EECS [6], eLEACH [7]	1 hop, variable transmission power	random nodes	random form
1-HOP GRID	HEED [8], BP [9], PC [10], EEPA [11]	1 hop, fixed transmission power	min/max distance from other CHs	quasi-circular
k-HOP	FLOC [13], EDC [14], k-random [15], ConID [16], UCCP [17], Max-Min [18], LNCA [19]	parameter k		
LOCATION BASED	Clique [2], GROUP [20], Multi-Res [21]	any possible	location-based	exact (squares, hexagons, etc.)
CENTRALIZED	HCR [12], ILP [22]		any possible	any possible

Fig. 1. Summary of cluster properties for some state of the art protocols.

2 Current practice for clustering algorithms

A wide variety of clustering algorithms has been developed with different properties. This section offers a high level survey of these approaches as well as previous works considering optimal clustering.

2.1 State of the art clustering protocols

This survey is not intended to be exhaustive or complete as it is impossible to do so in the space allowed. We have, however, identified five main families of protocols: random, 1-hop grid, k-hop, location-based and centralized clustering protocols. Their main properties and our taxonomy are summarized in Figure 1.

Random clustering. Many clustering protocols are improvements or modifications to LEACH [1], in which network nodes choose to be cluster heads based on a probability known a priori at all nodes. Self-elected cluster heads flood a cluster head role assignment message to their neighbors, which in turn identify and select the nearest cluster head. In the original LEACH protocol, the probability corresponds to the number of desired cluster heads in the network. Additional metrics such as remaining node energy [6, 7] can also be used to change the clustering properties. LEACH-like random-clustering algorithms build clusters with completely random sizes and shapes. Additionally, cluster heads are independent from one another and can be located anywhere in the network. The original LEACH assumes that it is always possible for a node to reach any cluster head in one hop, however, nodes are allowed to vary their transmission power.

1-hop grid clustering. Assuming full network connectivity is not reasonable in all scenarios, therefore multi-hop topologies need to be addressed. Two different families of protocols have evolved over time: 1-hop grid and k-hop fixed transmission power clustering algorithms. Representatives of the 1-hop grid clustering protocols are HEED [8], BP [9], Passive Clustering (PC) [10], or EEPA [11]. These protocols require the cluster head in any cluster to be able to communicate to its neighboring cluster heads in one hop, thus building a virtual grid. Consequently, they assume very dense networks. The shape of the resulting clusters is semi-circular and the size is bounded by the communication radius of the nodes. For these algorithms it is important to keep the number of clusters as low as possible and often the *optimal clustering* is defined as the one that minimizes the number of clusters while meeting the 1-hop grid communication requirement.

K-hop clustering. The second family of protocols, including FLOC [13], EDC [14] and others [15–17] extend the size of the clusters to multiple hops between cluster members and cluster heads, thus also eliminating the virtual grid of the 1-hop grid clustering (see above). Again, they first randomly assign cluster head roles to some nodes in the network and then "grow" clusters around them. In case a node cannot find a cluster head at most k hops away, it becomes a "forced" cluster head [15]. Others [16,18] use k-hop neighborhood information to optimize clusters and cluster heads: for example selecting the lowest ID as the cluster head. UUCP [17] uses optimization techniques from operations research to find a well-balanced cluster head. As with 1-hop grid algorithms, the number of clusters should be minimized, such that most of the clusters are exactly k-hops wide.

In LNCA [19] nodes first exchange information about their data readings, then, according to similarity of data, form k-hop clusters. As such, it is one of the rare efforts to match the size and shape of clusters to the gathered sensory data: nodes form clusters only if their data is similar and can be aggregated with little or no data loss. From the cluster form perspective the algorithm is a traditional k-hop clustering, but with random cluster sizes because of the data similarity requirement.

Location-based clustering. Geographic, or location-based clustering protocols have well defined cluster sizes and shapes, which are usually parameters. GROUP [20] builds a location-based grid with quadrants of tunable size. This grid is laid over the network and nodes next to the grid crossing points become cluster heads. Another geographic-based clustering approach is applied in [21] for multi-resolution in-network storage of data for WSNs. In this case a hash function is used to map the cluster head roles to network locations: the nearest nodes to those locations become cluster heads and store aggregated data for further reference. Our own clustering protocol Clique [2] is also geographic-based with variable cluster sizes. However, cluster head roles are not assigned but instead are decided on the fly as data packets are being forwarded to the base stations.

Centralized clustering. There are many clustering algorithms that require full network topology and/or remaining energy information to centrally compute optimal clusters (e.g. [12, 22]). At each round they disseminate the cluster information to all nodes. These protocols can clearly build any clusters with any properties. However, such approaches do not scale and do not consider fundamental network issues such as failures and asymmetric links.

Data aggregation and clustering. One major goal of clustering is to allow in-network pre-processing (aggregation or compression), assuming that cluster heads (and other intermediate nodes) collect multiple data packets and relay only one aggregated/compressed packet. [23] identifies three different aggregation techniques: tree aggregation, centralized pre-processing and gossiping. The first refers to the case in which data is processed and aggregated at each hop. Thus, the task of aggregation is not limited to the cluster head, but is spread over many nodes in the cluster. This is a great advantage especially in multi-hop clusters. The second refers to a LEACH-like clustering and aggregation scheme in which the data of the whole cluster is gathered on one central node (cluster head) and pre-processed there. If the cluster is multiple hops wide, however, this aggregation scheme has a greater communication overhead compared to a tree-based one. On the other hand, data processing itself is more precise, since all raw data readings are available. The third aggregation technique describes the case where no clusters are maintained: instead, nodes exchange (gossip) some of their data readings with other nodes, typically randomly. As this family of protocols is not related to clustering, we do not discuss it further.

2.2 Evaluation methodologies for clustering algorithms

Next we concentrate on *how* the above protocols were evaluated rather than the results they achieve. Crucial for the comparison of different experimental results are the evaluation environment and the selected metrics. Unfortunately, many works do not state which evaluation environment they use, reporting only that it is simulation. From the above surveyed protocols only one has been evaluated on a real testbed [13].

It is interesting to observe which evaluation metrics researchers apply to their clustering algorithms. Measuring energy expenditure or communication overhead is common, but not universal. This is unfortunate, because all clustering work is predicated on the fact that applying clustering reduces network energy expenditure. When energy expenditure is evaluated, sometimes it is considered after the clusters have been built while others include the overhead to build the clusters. Still others use network lifetime, usually defined as the time of first node death. Nearly all of the protocols have been evaluated in terms of the number of clusters or cluster heads, interpreting a low number of clusters as a *good* result. The underlying assumption for this is that when the cluster size is bound to k-hop communication, a lower number of clusters means optimal clustering. While this may be true if the right k value is used, there is no investigation of how to find the right k. Furthermore, if the protocol does not restrict the size of the clusters, a low number of clusters may result in very high in-cluster communication overhead due to the increase in single cluster size.

One good evaluation criteria is the standard deviation of the number of nodes in a cluster. This shows clearly the balance of the cluster sizes, which ensures uniform data aggregation throughout the network. Unfortunately, however, this metric is not considered for all protocols we studied.

In general, there are two evaluation scenarios which need to be covered by the researchers to show the complete behavior and performance of a new protocol. We call the first *fixed network evaluation* and the second *scalability analysis.* The fixed network evaluation is what most researchers do in their comparative studies. They *fix* the network, the application, the data traffic, etc. and compare several protocols in terms of network lifetime, incurred overhead for clustering and routing, energy expenditure etc. In fact, such an evaluation is meaningful only in a comparative analysis. In contrast, measuring some of these properties for an isolated protocol is often meaningless. For example, it is impossible to evaluate whether the reported network lifetime for protocol X is sufficient or not.

Instead, scalability analysis allows independent protocol analysis as it is intended to show the behavior of the new protocol with various network settings such as network size, number of nodes, data traffic, etc. Here, the results of a comparative analysis can actually be misleading. For example, the clustering overhead of some protocol may skyrocket with increasing node density while a new protocol may show slightly lower overhead. This does not mean that the new protocol is scalable and performs well in high density scenarios. It may, instead, mean that *both* protocols are not scalable. Therefore, analysis is needed where the new protocol is evaluated in isolation. Of course, it is often appropriate to offer a scalability analysis for both the presented work as well as its competitors. In this case the evaluation becomes again comparative, however the methodology is different.

Unfortunately, none of the works presented in our survey show both types of evaluations. Admittedly, this is likely due in part to page limitations, but the required time and effort also contribute. Some of the works opt for the scalability analysis (e.g. [13,15,16]), and others for comparative evaluation (e.g. [2,8,9,17]). We also note the quality of the evaluation. For example, some comparative analyses do not use earlier protocols for comparison, but instead compare against their own trivial clustering implementation. Alternately, comparing a clustering protocol against scenarios with no clustering was reasonable in the early years of clustering research [1, 10], however can no longer be considered a valid comparison [12].

While these criticisms imply a clear need for benchmarks for comparison among clustering approaches in WSNs, this paper concentrates on identifying the *optimal* clustering clustering scheme. Together with communication benchmarks, the optimal clustering will simplify comparison among different algorithms and protocols and even enable comparison at a very high level, e.g. cluster shapes and sizes.

2.3 Related efforts in clustering analysis

This paper is not the first to step back and analytically evaluate clustering techniques. In a very recent effort [24], the author addresses the question: given a

PARAMETER		DESCRIPTION
network	N	number of nodes
	A	size of network (meters)
	r_c	communication radius of nodes (meters)
	M	number of base stations
clustering	C	size of cluster (meters)
	P_{CH}	position of cluster head [center, random, station]
	I_N	in-network processing [tree, center]

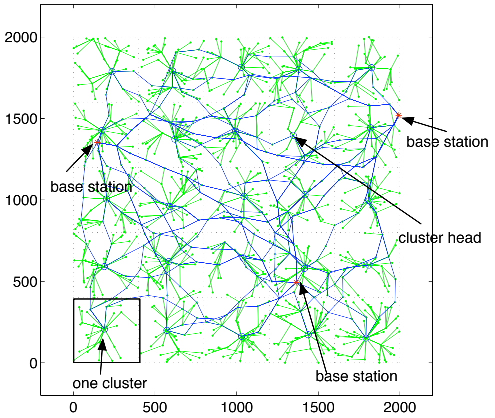


Fig. 2. Network and clustering parameters with a sample network of 2000 nodes.

network with N uniformly spread sensors, how big is the optimal cluster measured in the number of sensors? In this work the network model assumes that the network can be divided into cells with each cell containing a single sensor with a very high probability. Additionally, sensors can communicate to all their adjacent sensors. The cluster heads are always in the center of the cluster and have more powerful radios to be able to communicate to all adjacent cluster heads. In this model, the question of the optimal size of a cluster is reduced to the calculation of the number of transmissions required to reach the cluster head and the single base station. The author's answer to the above question is: *In a network with $A \times A$ cells, where each cluster is $x \times x$ cells big, the optimal x is as close to $\sqrt[3]{2N}$ as possible and divides A* [24]. For example, for a very big network, e.g. 1156 nodes (34x34 cells) the optimal cluster size is 12 (or 6 hop cluster radius). For a small network, e.g. with 256 nodes (16x16 cells) the optimal cluster size is 4 (or 2 hop cluster radius).

Other works report different results. The analysis described in [19] concludes that a cluster radius of 2 hops is optimal for any practical network of 300-2000 nodes. However, the authors use multi-hop routing through normal sensors to reach the base station instead of cluster heads only. The difference between the results can be attributed to the difference in the chosen network models. In this work we extend the network model, allowing many more parameters. In subsequent sections, we discuss in detail the discrepancies between ours and the above presented results.

3 Defining the optimal cluster

Our first step is to extend the network models used by [19, 24] to incorporate multiple network and cluster parameters, as summarized in Figure 2, and then to analyze optimal cluster sizes. We define the WSN to be a flat network with N nodes, uniformly and randomly spread over a square area with size A. We assume that clusters are also squares with some size C. Nodes can communicate to all their neighbors, defined as those nodes whose distance is less than some communication radius r_c. Energy is spent when a node sends or receives a packet. We do not use a specific energy model to calculate the exact energy expenditure, but instead always show the number of sent/received messages (ETX+ERX). As we assume a broadcast environment neighbors receive messages even if they are not destined to them.

In every cluster there is a single cluster head, reachable in k hops by all cluster members. The position of the cluster head inside the cluster is an input parameter, P_{CH}, that can be set to: near the center of the cluster, near the base stations (in case of multiple base stations the minimum distance sum to all of them is used as metric), or random. Each node gathers sensory data and sends it first to the cluster head, then the cluster head aggregates the received packets and sends a single packet to all base stations. Multi-hop routing through all sensor nodes (cluster heads and cluster members) are used for both aggregated and non-aggregated packets. There are M base stations in the network, randomly selected among all nodes, therefore they have no special properties such as increased battery or communication range. In-network processing (aggregation, compression) is either tree-based or centralized at the cluster head (see also Section 2.1 or [23]) (parameter $I_N = \{tree, CH\}$). The network parameters we use are summarized in Figure 2.

We evaluate the performance of a clustering scheme with given parameters $N, A, r_c, C, P_{CH}, M, I_N$ in terms of the number of received/sent packets for routing the sensory data from the sensors through the cluster heads to the base stations. We define:

Definition 1. *the optimal clustering scenario is the 3-tuple $\{C, P_{CH}, I_N\}$ which incurs the minimum communication overhead for the network $\{A, N, r_c, M\}$.*

Definition 2. *the communication overhead of a network $\{A, N, r_c, M\}$ with clustering scenario $\{C, P_{CH}, I_N\}$ is the sum of sent packets and received packets for all nodes in the network for one round of data reporting. In one round of data reporting each node sends exactly one packet to its cluster head and the cluster heads send exactly one packet to all base stations M. Multi-hop routing is used for all transmissions.*

Note that the network model and clustering scenarios we define here are more general and sophisticated than those previously proposed [19,24]. We allow more parameters (node density, communication radius, multiple base stations) and different aggregation schemes (tree-based, centralized).

The questions we address in the next section are:

1. Do general rules exist for optimal clusters? For example, do 2-hop clusters perform the best for all network sizes, independent of the number of nodes, network area or cluster head positions?
2. If there are no rules for all parameters, what are the rules of thumb for selecting the cluster parameters $\{C, P_{CH}, M\}$ for some given network $\{N, A, r_c\}$?

3. Do the above results change when a different in-network aggregation scheme is used, e.g., tree-based vs. centralized?

4 Identifying the optimal cluster

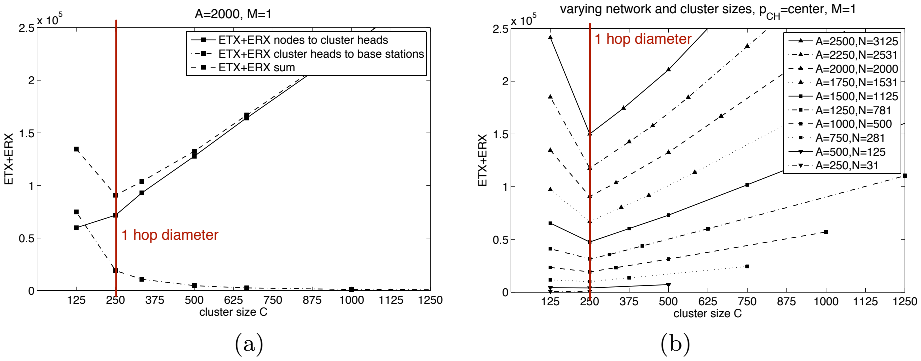
Our analysis considers the effect of key clustering parameters on the communication overhead. We follow an experimental approach for two reasons. First, it is hard if not impossible to derive generally valid formulas for the network communication overhead that consider all parameters, especially random topologies. Such a theoretical approach has been previously done [19, 24], however several required, simplifying assumptions make the results difficult to apply in practice. Further, we extend the network models of these works to accommodate different densities and fully random topologies. Our second motivation is to make our results immediately applicable: a WSN practitioner can select the most relevant scenarios from our experiments and directly derive the optimal clustering parameters.

We performed our simulations in MATLAB; the source code and the raw data are publicly available at `www.sensorlab.inf.usi.ch`. Figure 2 shows a sample network. Nodes are spread randomly through the network field. The network area is divided into equal-size clusters, and all results are presented for a variety of cluster sizes that allow the area to be precisely divided into such equal-size clusters. The shortest distance in terms of ETX, expected number of transmissions, is computed between each node and its corresponding cluster head and between all cluster heads and the base stations. The energy expenditure is calculated as the sum of ETX and ERX (expected number of receivers) for one round of data gathering. Overhearing costs are included in ERX. Cluster formation overhead is ignored, because we do not assume any particular clustering scheme and thus cannot calculate this overhead. Each of the reported experiments is the mean of 100 independent random topologies with random nodes selected as base stations.

Our analysis addresses the energy expenditure of different clustering schemes by exploring the parameters A, C, M, P_{CH}, I_N, N. Our goal is always to identify the optimal cluster size for scenarios, first studying the optimal cluster size for a scenario with no intra-cluster aggregation ($I_N = center$), the cluster head near the center of the cluster ($P_{CH} = center$), and constant node density ($N/A^2 = CONST \approx 500 nodes/km^2$). We then vary the number of base stations, the position of the cluster head, the use of in-network processing, and finally the node density.

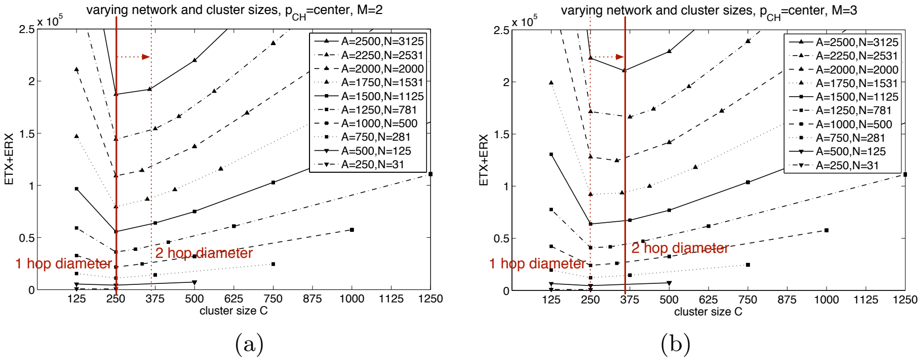
4.1 Cluster size C.

We first consider a single setting whose key parameters are described in Figure 3. To understand the components of the communication cost, Figure 3(a) separates the cost into intra and inter cluster communication. Intuitively, as cluster size



(a) (b)

Fig. 3. The minimal communication overhead is incurred with 1-hop wide clusters. (a) Energy expenditure for $A = 2000m, M = 1, r_c = 150m, P_{CH} = center, I_N = center, N = const$ separated into intra and inter-cluster communication overhead. (b) Total communication overhead for varying network sizes (different lines).



(a) (b)

Fig. 4. The minimal incurred communication overhead shifts from 1-hop to 2-hop wide clusters for growing number of sinks in the network. Energy expenditure for $r_c = 150m, P_{CH} = center, I_N = center, N = const$ and (a) M = 2; (b) M = 3.

grows, communication inside the clusters also grows as data needs to be forwarded multi-hops to the cluster head. At the same time communication from cluster heads to the base station drops significantly, since fewer cluster heads are present. The minimum point on the sum of the two lines shows the optimal cluster to be approximately 250, or 1-hop clusters because in our scenario $r_c = 150m$.

Figure 3(b) confirms the optimality of 1-hop clustering for a wide range of network sizes with 30–3000 nodes, but always constant mean density (constant ratio of nodes per deployment area).

This result stands in contrast to those previously reported in the literature [19, 24]. Differences from [24] follow from different network models. Notably, [24] does not take into account routing from cluster heads to base stations and assumes a very regular topology with a single node able to communicate to

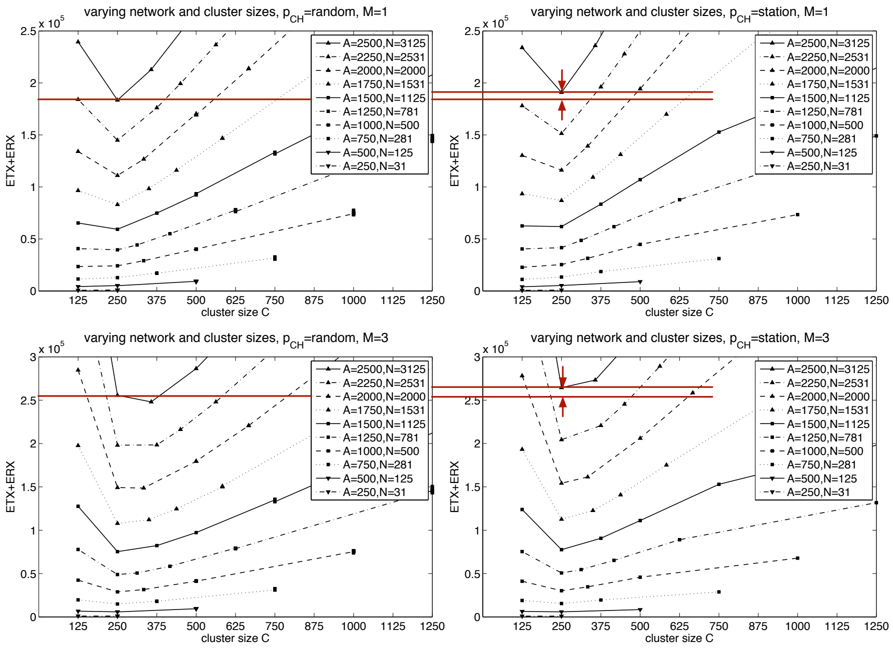
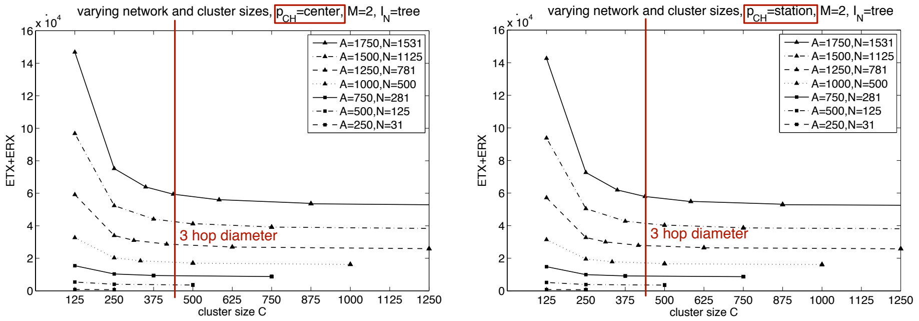


Fig. 5. Energy expenditure with random cluster heads is slightly lower than with base station oriented ones. Energy expenditure for $r_c = 150m, I_N = center, N = const$ and (a) $P_{CH} = random, M = 1$; (b) $P_{CH} = random, M = 3$; (a) $P_{CH} = station, M = 1$; (b) $P_{CH} = station, M = 3$.

exactly four neighbors. Instead, our random topologies allow different node densities in different parts of the network plus we include routing overhead to reach the base stations. On the other hand, we believe the difference from [19] is due to the fundamental difference between an analytical analysis and experimentation. While they provide general formulas with parameters for the network and cluster sizes, their analysis relies heavily on the same virtual grid topology as [24] and makes many assumptions and generalizations about energy expenditure.

4.2 Number of base stations M

Another key parameter to evaluate is the clustering behavior with different number of base stations collecting the results. We keep the same parameters as in the previous section and Figure 3(b), but extend the number of base stations to $M = 2$ in Figure 4(a) and $M = 3$ in Figure 4(b). Random nodes are selected to be base stations in each experimental run. As the number of base stations increases, inter cluster communication grows as well. In terms of the independent overheads represented in Figure 3(a), the intersection moves to the right, implying larger optimal clusters. This trend is visible in Figure 4(a), where for large networks (e.g., the top line) the energy expenditure is nearly the same for 1 and 2 hop clusters. However, with three base stations and large networks the



(a) (b)

Fig. 6. With tree-based aggregation the optimal cluster size lies at 3-4 hop wide clusters. The position of the cluster heads is irrelevant. Energy expenditure for $r_c = 150m, I_N = tree, N = const, M = 2$ and (a) $P_{CH} = center$; (b) $P_{CH} = station$.

optimal cluster size is 2 hops (cluster size ≈ 375m with $r_c = 150m$). We expect this trend to continue for very large networks of tens of thousands of nodes and plan to verify this in the future.

4.3 Position of the cluster head P_{CH}

Next, we explore the effect of the cluster head position inside the cluster. While the previous experiments shown in Figures 3 and 4 placed the cluster head close to the center of the cluster, here we allow it to be random (Figure 5(a-b)) or to be closest to the base station (Figure 5(c-d)). The position of the cluster head is important for two reasons. First, it affects the load balance, and therefore energy consumption, inside the cluster for routing and data aggregation. Second, the routing overhead inside the cluster changes with the head placement. Based on our analysis, we make two key observations: first, the optimal cluster size is not affected by P_{CH}. Second, the cluster head position does affect the total energy spent in the network. Specifically, when the cluster head is at the cluster center (Figure 3), it requires 15% less energy than a random placement due to intra cluster routing costs. Notably, with head placement closest to the base stations, the clustering scenario does not perform better than the other options (Figure 5(c-d)). Even though the routing overhead between the cluster heads and the base stations is minimized, this savings is outweighed by the increased routing inside the clusters.

4.4 In-network processing I_N

Next we consider the possibility to process the data inside a cluster, as it is being forwarded to the cluster head. While this ability typically depends on the application and cannot be changed simply to reduce overhead, in applications where either option is feasible our analysis in Figure 6 shows that tree-based

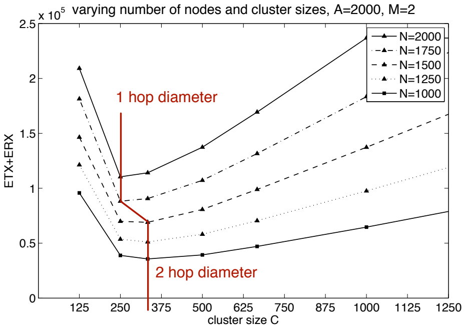


Fig. 7. Larger clusters are more energy-efficient when node density is low. Energy expenditure for $A = 2000, r_c = 150m, P_{CH} = center, I_N = center, M = 2$

aggregation is preferable. Intuitively, the total energy expenditure decreases with increasing cluster sizes because the data aggregation rate grows and data traffic decreases. However, for very large clusters the gains are rather insignificant. Consequently, preference should be given to 3-4 hop clusters since they have simultaneously low energy expenditure and lower data aggregation rates.

Interestingly, the tree-based aggregation removes the importance of the cluster head position, as seen by comparing Figures 6(a) and 6(b). The energy expenditure is the same because the effect of in-cluster routing was also eliminated.

4.5 Node density N/A^2

In our final experiment we vary the node density with a fixed network size. Figure 7 shows a clear trend that lower densities (≈250-375 $nodes/km^2$) result in larger optimal clusters. For higher densities (≈400-500 $nodes/km^2$) the optimal cluster size is again 1 hop. This is because low node densities lower the total intra cluster communication overhead, giving the inter cluster communication more significance.

5 Conclusions and Future Work

Although it is intuitive that clustering has the potential to reduce energy consumption, our analysis into optimal clusters reveals some general guidelines for WSN practitioners.

First, **1-hop clustering performs best for a large spectrum of different network sizes, node densities and number of base stations**. For very large networks (more than 1000 nodes), multiple base stations (more than three) or very low densities (less than 400 $nodes/km^2$) 2-hop clustering performs better, although not significantly. Additionally, **the optimal cluster head position is the center of the cluster**. In comparison to random locations or those closest to the base station, it spends approximately 15% less energy. **For tree-based**

aggregation, 3 to 4-hop clustering performs best in terms of energy expenditure and data aggregation rate. Here the position of the cluster head inside the cluster is not important.

Of those protocols surveyed in Section 2, optimal clustering is achievable by k-hop clustering algorithms. Additionally, location-based algorithms are also promising since they can accommodate any hop diameters. Nevertheless, their parameterization may be difficult, since the network topology must be known a priori to calculate the optimal cluster size.

While our analysis identifies how the optimal cluster looks, it does not answer the question of how to find and build this clustering with as little overhead as possible, or how to spread the communication and computation load among all nodes in the network. This opens avenues for future protocol development.

Our near term research goals include extending our analysis in several directions. First, we will analyze the scalability of clustering scenarios up to tens of thousands of nodes. Second, we will explore the communication overhead of the individual nodes to understand how the load can be optimally spread. Finally, we will conduct more realistic experiments using a sophisticated network simulator.

References

1. Rabiner-Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: Energy-Efficient Communication Protocol for Wireless Microsensor Networks. In: Proc. of the 33rd Hawaii Int. Conf. on System Sciences, Washington DC, USA (2000)
2. Förster, A., Murphy, A.L.: CLIQUE: Role-free clustering with Q-learning for Wireless Sensor Networks. In: Proc. of the 29th Int. Conf. on Dist. Computing (ICDCS), Montreal, Canada (2009)
3. Raman, B., Chebrolu, K.: Censor Networks: A Critique of "Sensor Networks" from a Systems Perspective. ACM SIGCOMM Comp. Comm. Review **38**(3) (2008)
4. Kurkowski, S., Camp, T., Colagrosso, M.: Manet simulation studies: the incredibles. SIGMOBILE Mobile Comp. and Comm. Revue **9**(4) (2005) 50–61
5. Heidemann, J., Bulusu, N., Elson, J., Intanagonwiwat, C., Lan, K.C., Xu, Y., Ye, W., Estrin, D., Govindan, R.: Effects of detail in wireless network simulation. In: Proc. of Society for Comp. Simulation Comm. Networks and Distr. Systems Modeling and Simulation Conf. (CNDS). (2001)
6. Ye, M., Li, C., Chen, G., Wu, J.: Eecs: an energy efficient clustering scheme in wireless sensor networks. Proc. of the 24th IEEE Int. Conf. on Performance, Computing, and Communications (April 2005) 535–540
7. Jang, K.Y., Kim, K.T., Youn, H.Y.: An energy efficient routing scheme for wireless sensor networks. In: Proc. of the Int. Conf. on Computational Science and its Applications. (2007) 399–404
8. Younis, O., Fahmy, S.: Heed: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. IEEE Trans. on Mob. Comp. **3**(4) (2004)
9. Anker, T., Bickson, D., Dolev, D., Hod, B.: Efficient clustering for improving network performance in wireless sensor networks. Proceedings of the 5th Eur. Conf. on Wireless Sensor Networks (2008)

10. Gerla, M., Kwon, T., Pei, G.: On demand routing in large ad hoc wireless networks with passive clustering. In: Proceedings of IEEE Wireless Comm. and Netw. Conf. (WCNC), Chicago, USA (2000) 100–105
11. Yu, M., Leung, K., Malvankar, A.: A dynamic clustering and energy efficient routing technique for sensor networks. IEEE Trans. on wireless comm. **6**(4) (2007)
12. Al-Karaki, J.N., Ul-Mustafa, R., Kamal, A.E.: Data aggregation in wireless sensor networks - exact and approximate algorithms. In: Proc. of the Works. on High Performance Switching and Routing, Phoenix, AZ (2004)
13. Demirbas, M., Arora, A., Mittal, V., Kulathumani, V.: Design and analysis of a fast local clustering service for wireless sensor networks. In: Proc. of the 1st Int. Conf. on Broadband Wireless Networking (BroadNets). (2004) 700–709
14. Chen, Q., Ma, J., Zhu, Y., Zhang, D., Ni, L.: An energy-efficient k-hop clustering framework for wireless sensor networks. In: Proc. of the 4th Eur. Conf. on Wireless Sensor Networks (EWSN). (2007)
15. Bandyopadhyay, S., Coyle, E.: An energy efficient hierarchical clustering algorithm for wireless sensor networks. In: Proc. of the Annual Joint Conf. of the IEEE Comp. and Comm. Societies (INFOCOM). Volume 3. (March 2003) 1713 – 1723
16. Nocetti, F., Gonzalez, J., Stojmenovic, I.: Connectivity based k-hop clustering in wireless networks. Telecommunications Systems **22**(1-4) (2003) 205–220
17. Aslam, N., Phillips, W., Robertson, W.: A unified clustering and communication protocol for wireless sensor networks. IAENG Int. J. of Comp. Sc. **35**(3) (2008)
18. Amis, A., Prakash, R., Vuong, T., Huynh, D.: Max-min d-cluster formation in wireless ad hoc networks. In: Proc. of the 19th Annual Joint Conf. of the IEEE Computer and Communications Societies. Volume 1. (2000) 32–41
19. Xia, D., Vlajic, N.: Near-optimal node clustering in wireless sensor networks for environment monitoring. Proc. of the 21st Int. Conf. on Advanced Networking and Applications (2007)
20. Yu, L., Wang, N., Zhang, W., Zheng, C.: Group: A grid-clustering routing protocol for wireless sensor networks. In: Proc. of the Int. Conf. on Wireless Comm., Networking and Mobile Computing (WiCOM), Wuhan, China (December 2006)
21. Ganesan, D., Greenstein, B., Estrin, D., Heidemann, J., Govindan, R.: Multiresolution storage and search in sensor networks. ACM Trans. on Storage **1**(3) (2005)
22. Matin, A.W., Hussain, S.: Intelligent hierarchical cluster-based routing. In: Proc. of the Int. Works. on Mobility and Scalability in Wireless Sensor Networks (MSWSN), San Francisco, CA (2006)
23. Chitnis, L., Dobra, A., Ranka, S.: Aggregation methods for large-scale sensor networks. ACM Trans. on Sensor Networks **4**(2) (March 2008)
24. Wang, D.: An energy-efficient clusterhead assignment scheme for hierarchical wireless sensor networks. Int. Journal of Wireless Inf. Networks **15**(2) (2008) 61–71