Investigation of Information Dissemination Design Criteria in Large-scale Network Environments

Konstantinos Oikonomou
Dept. of Informatics
Ionian University
Corfu, Greece
Email: okon@ionio.gr

Dimitrios Kogias, Leonidas Tzevelekas, Ioannis Stavrakakis
Dept. of Informatics and Telecommunications
National and Kapodistrian University of Athens
Athens, Greece
Email: {dimkog, ltzev, ioannis}@di.uoa.gr

***Abstract*—The design of efficient *information dissemination* mechanism is a challenging problem in large-scale network with respect to the *number of messages* and *termination time*. In this paper, *advertisement* and *searching* – the two basic ingredients of information dissemination – are investigated and certain *criteria* are proposed with respect to the correctness, promptness and fairness of the approach. Based on the *complementarity* of both advertisement and searching, the aforementioned criteria can be satisfied under certain conditions, which form the baseline of design principles for efficient information dissemination, as analytically – also using numerical results – is investigated here.**

***Keywords*-Information Dissemination; Dominating Sets; Large-scale Networks;**

I. Introduction

Information dissemination in modern network environments – like peer-to-peer (P2P) networks, autonomic networks, ad hoc networks, sensor networks, etc. – that are typically large-scale, is a challenging problem, with respect to the *number of messages* and *termination time* as the number of nodes increases. Information dissemination mechanisms are part of many proposed and implemented algorithms and network protocols. For example, in routing it is common for appropriate messages to be send in the network advertising or searching for a certain destination (and tracking the route towards the particular destination). Both *advertising* and *searching* may be seen as *complementary* ingredients of a particular information dissemination approach. Complementary in the sense an *intensive* advertisement would cover a larger network area (than less intensive), thus allowing for less intensive searching. The study of this particular complementary nature of the dissemination information ingredients, is basically the focus of this work.

A first obvious but costly approach to information dissemination (for both the advertisement and the searching phasse) is *traditional flooding*, [1]. Under traditional flooding all network nodes receive a certain message which eventually traverses all network links resulting in a number of messages of the order of N^2, while the optimal is $N-1$ messages for a network of N nodes (e.g., forwarding messages over the branches of an overlay spanning tree). *Termination time*, is significantly small and upper bounded by the *network diameter*, typically of the order of $\log(N)$.

If traditional flooding is used for the advertisement, it is interesting to note that searching is obsolete since the disseminated information has reached all network nodes. Similarly, if it is a priori known that searching will employ flooding, there is no need at all for advertisement, since all network nodes will eventually be reached when searching for the particular piece of information. Another approach is to use *probabilistic flooding*, [2], [3], [4], [5], that probabilistic ensures that all network nodes will receive the particular message with high probability (e.g., that means that some nodes may not receive the particular message) and the number of message will be significantly reduced (of the order of $\log(N)$, for appropriate values of the forwarding probability, [4]). Termination time is increased but not that significantly (remains of the order of $\log(N)$.

Even if probabilistic flooding is employed, still a large number of messages is required in a way that it is prohibitive in large-scale network to allow global network outreach information dissemination policies. For this reason, variations of flooding have been proposed. For example, regarding searching in the Gnutella P2P system, [6], a TTL (Time-To-Live) value L is used to restrict message flooding to a small number of hops around the node that has initiated searching (to be called hereafter the *initiator node*). This approach, referred to hereafter as *L-flooding*, may be scalable for small values of L but at the same time it significantly reduces the probability of locating the requested node(s) of interest in large P2P networks (i.e., large values for N).

Other approaches, like *random walks*, e.g. [7], have been proposed to reduce the total number of messages by sending a limited number of special messages (agents) in the network. Each of them follows its own path by choosing randomly the next hop node. Messages terminate their walk either after some time (e.g., TTL expiration) or after checking with the initiator node and learning that the node of interest has already been discovered by another message, or a combination of both. Hybrid probabilistic schemes (e.g., a local flooding process initiated after a random walk) have also been proposed and analyzed, [8], as well as other schemes that adapt the employed TTL values in a probabilistic manner, [9]. Another modification, [10], allows for network nodes to forward messages to their neighbors in a random manner, thus significantly reducing the number of messages in the network. Many other

works have been published proposing the selective forwarding of a certain message in the network, e.g., [11], [12], [13], [14], [15].

Recently, more elaborate ideas on random walkers have been introduced attempting to fill the gap between traditional flooding and single random walkers. For example, the idea of random walkers with jumps, [16], has been introduced in order to move random walkers to "different" network areas and avoid the problem of "oversampling" a certain network area. The idea of multiple random walkers, [17], has also been introduced allowing a random walker to "split" according to a certain splitting policy and move towards different network directions. These, approaches, apart from the fact that it has been shown to improve both coverage and termination time when compared to a single random walker, still employ a probabilistic mechanism, as opposed to the deterministic mechanism employed by traditional flooding. Consequently, any argument with respect to their performance is made "with high probability" being based – basically – on averaged values.

Some approaches (e.g., flooding) result in an increased number of messages but they always (ignoring the possibility of system faults) return a *correct* reply (i.e., the initiator node location is always retrieved if it is available in the network, otherwise a negative reply is returned) within a certain time limit (which may be increased depending on the network diameter). Other approaches (e.g., Gnutella) do not always return a correct answer (i.e., the particular piece of information might be in the network but the searching mechanism may fail to locate it), even though replies are sent within certain time limits and the number of messages is bounded. In other cases (e.g., random walkers), the time required until a reply is received may become significantly high in order to get a correct reply. Furthermore, this particular treatment should be applied to all network nodes in a suitable manner.

An important aspect of advertisement and searching is their *complementary* nature. In particular, this complementarity property reveals a tradeoff in the *intensity* of the two phases and by selecting the intensity and the mechanism for one of them, it is possible to shape expectations for the other one, which is analytically investigated in this paper. Given a searching mechanism and intensity, the (complementary) advertisement can be designed and parameterized so that the combined (complementary) dissemination information process meats important criteria such as *correctness* (i.e., reach the appropriate set of nodes), *promptness* (i.e., meet certain time limits) and *fairness* (i.e., equally apply all criteria to all network users).

In order to study the aforementioned complementarity, a certain searching policy is assumed here – similar to L-flooding as it is the case in Gnutella, [6] – such that any initiator network node that is searching for a particular peace of information (e.g., a service, a file, a route) sends messages to all network nodes that are located at most L hops away. L is regarded as a *measure of the intensity* of the particular searching policy and is an important factor regarding of the intensity of the (complementary) advertising phase, as mentioned before. Actually, by defining the form of searching, this gives further flexibility on studying advertisement.

Subsequently, in order to satisfy all the aforementioned criteria for the given searching process and intensity, the advertisement process should be able to disseminate the particular piece of information of interest at most L hops away from any network node. The latter property, to be referred to hereafter as the L-*property*, can be satisfied by many different algorithms (e.g., flooding-based, random walks). As it is expected and as it is shown in the subsequent analysis included in this paper, a suitable selection of L may allow for reducing the total number of messages.

On the other hand, as it is shown here, when *minimization* of the number of messages sent during the advertisement process is required, this turns out to be a *dominating set* problem, [18], that is known to be NP-hard and requires global network knowledge. Consequently, it is not a suitable approach for network environments of normally large number of nodes, due to the introduced high overhead both in time and number of messages.

II. Problem Definition

The aim in this section is to exploit the complementary nature of both advertisement and searching by setting certain requirements for the former phase derived from a careful consideration of the latter phase. Obviously, from the latter phase it is required to satisfy both the correctness criterion (i.e., to reach a certain set of nodes) and the promptness criterion (i.e., certain time limits to be satisfied). Assume that searching employs a L-flooding algorithm for searching in the area of nodes located at most L hops away from any initiator node u. Given that the searching policy is only allowed to look for the required information in the area of L hops away from any network node, the particular information *must be available* in the particular area, otherwise the correctness criterion will not be satisfied (further discussion on L is provided in Section IV). Subsequently, it is the role of the advertisement policy to satisfy this particular criterion. Before going into the details, some useful definitions are given next.

Let the undirected graph $G(V, E)$ represent a network with a certain set of nodes V and a set of bidirectional edges E among nodes. Let $|X|$ denote the number of elements or size of a particular set X. For the rest the number of nodes $|V|$ in the network will also be noted as N ($N = |V|$).

Assume that A piece of information to be disseminated is initially located at a certain node s, which will be referred to hereafter as the *information node*. The objective of the advertisement process is to inform a subset of the network nodes $V_a \subseteq V$ about the location of the information node s. Depending on the particular case, it may be required the advertising process to reach all network nodes ($V_a = V$) or only a small portion of them ($|V_a| < |V|$).

Since information about the information node is initially available at node s, the advertisement process is assumed to be initiated by the particular node and sent to the network. Based on the particular algorithm employed by the advertisement

process, a certain *advertising network* is created consisting of those nodes that have received information about the information node s (i.e., set V_a) and those links over which the particular messages were forwarded (i.e., set E_a). Let $A(V_a, E_a, s)$ denote the particular advertising network created for the network represented by graph $G(V, E)$, when the particular piece of information of interest is initially located at node s, for the particular advertisement policy. The number of links $|E_a|$ has an important meaning since they correspond to a lower bound of the number of messages sent in the network during advertisement and twice this number (i.e., $2|E_a|$) to an upper bound. In asymptotic terms, the number of messages are of the order of $\Theta(|E_a|)$. For the rest and for convenience, $|E_a|$ will be assumed to be the number of messages of the corresponding advertising policy.

Depending on the particular advertisement policy (actually, on the particular algorithm employed), the resulting advertising network $A(V_a, E_a, s)$ may be different. For example, if traditional flooding is used, then the resulting networks will be identical to $G(V, E)$ since all links and nodes of the network will be traversed by the flooded messages ($V_a \equiv V$ and $E_a \equiv E$). If probabilistic flooding is used, [4], then all network nodes (i.e., $V_a \equiv V$) will be reached *with high probability*, [4], but a smaller number of links will be traversed (i.e., $E_a \subset E$), for suitable values of the forwarding probability, [4].

Clearly, for the particular searching policy considered in the beginning of this section, the requirement is for a(n) (complementary) advertisement policy such that the resulting advertisement network satisfies the L-property formally defined next.

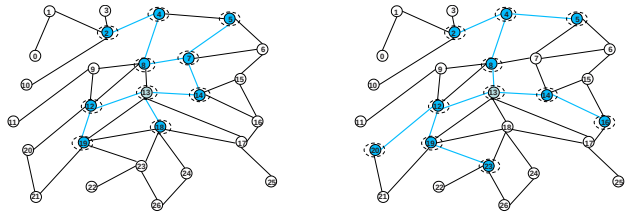
Definition 1: The L-property: For a particular network $G(V, E)$, for which the information of interest is located at the information node s, the advertising network $A(V_a, E_a, s)$, for a certain advertisement policy, satisfies the L-property, iff for all nodes $u \in V$ there exists at least one node $v \in V_a$, such that $d_{u,v} \leq L$.

The following section focuses on those advertising networks that satisfy the L-property and most important, on the complementary nature with respect to searching.

III. The L-Network

For a given network $G(V, E)$ and a certain information node s, it may be possible to create more than one advertising networks $A(V_a, E_a, s)$ that satisfy the L-property. Take for example the network depicted in Figure 1.a and Figure 1.b, where the information of interest is located at information node 13. For both cases an advertising network is also depicted and it is easy to see that both advertising networks satisfy the 2-property.

One possible way to create an advertising network that satisfies the L-property is to allow, for example, a random walker to node in the network for a sufficiently long time period such that the *trace* of nodes through which the random walker moved in the network is at most L hops away from an network node. Probabilistic flooding, [4], may also be used as well as other approaches already proposed in the literature (see Section I).



a. $|E_a| = 9$. b. $|E_a| = 10$.

Fig. 1. Example $G(V, E)$ network where the information of interest is located at the information node 13 and the corresponding advertising network $A(V_a, E_a, 13)$. The set of nodes V_a and the set of edges E_a are colored with dark grey.

Clearly, the advertising network depicted in Figure 1.a is preferable to the one depicted in Figure 1.b since the number of links is smaller. Smaller number of links for the advertising network means fewer messages required to create it and since both advertising networks satisfy the 2-property, the preference is one the one requiring fewer messages to be created. The smaller number of messages required to create an advertising network that satisfies the L-property, for a certain value of L, is actually the lower bound of the number of messages achievable under any advertising algorithm that aims to satisfy the L-property for a particular network and the searching policy considered in Section II, and provides for a suitable comparison basis among different approaches (e.g., flooding, probabilistic flooding, random walker, etc.).

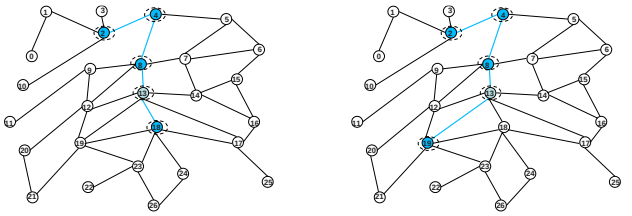
As it was presented so far, both criteria of correctness and promptness are satisfied when the advertisement network satisfies the L-property, for some value of L. Advertisement is responsible to create such an advertisement network. However, the minimization of the messages in the network should also be considered. Therefore, apart from creating an advertisement network satisfying the L-property, advertisement has to use the smallest possible number of messages. Eventually, the advertising network will be a L-network as defined next.

Definition 2: The L-network: An advertising network $A(V_a, E_a, s)$ for a particular network $G(V, E)$ that satisfies the L-property, is an L-network iff the number of links $|E_a|$ is the minimum among all advertising network $A(V_a, E_a, s)$ that satisfy the L-property for the particular network $G(V, E)$.

There can be more than one L-network for each particular case, as it is depicted in Figure 2, where both advertising networks are 2-networks (note that $|E_a| = 4$ for each case) but consist of different set of nodes and different set of links. For any different L-network it is clear that the number of links is the same and let E_L denote the particular (minimum) number of links.

An important and easily proved property of a L-network is given by the following lemma.

Lemma 1: A L-network is a tree. In addition, $|V_a| = |E_a| + 1 = E_L + 1$.



a. b.

Fig. 2. Different forms of a 2-network.

Proof: A L-network is an advertising network so it is a connected network since the advertisement process is initiated by the information node and consists of messages forwarded hop by hop according to the employed advertisement algorithm.

Since the L-network is by definition connected, and given that the requirement is the minimization of the number of links, it is clear that it does not contain any *cycle* in its topology (if it did, it could be removed and therefore, it would not be a L-network in first place). By definition, a network topology that is connected and contains no cycles is a tree.

Expression $|V_a| = |E_a| + 1 = E_L + 1$, is a direct result from the fact that a L-network is a tree. ■

Any algorithm that creates a L-network requires global information (i.e., knowledge of $G(V, E)$) and no polynomial time, i.e., any algorithm that creates a L-network is a NP-hard algorithm. This is easily derived from the fact that problem of creating any L-network is actually identical to a *dominating set* problem, [18], that is NP-hard.

The NP-hard nature of the problem and the requirement for global information, makes any attempt to propose any such algorithm for the advertisement process unsuccessful, when the particular network environment of large number of nodes is considered. On the other hand, any L-network, provides the minimum number of messages that are required for the creation of a certain advertising network satisfying the L-property, and it is another means for the evaluation of any future proposal for advertisement. Definitely, an important factor in this consideration is the suitable value of L, investigated in the following section.

IV. On the Appropriate Value of L

Let $K \subseteq V$ be the particular set of nodes that will, eventually, initiate a searching phase. For any node $u \in V_a$ (i.e., nodes that have received information about the information node location during advertisement phase) it is clear that no searching is required and the information is retrieved immediately. Consequently, the focus is on those nodes $J = K \setminus K \cap V_a)$. Assume for simplicity that these nodes will simultaneously start searching after termination of advertisement.

For any node $u \in J$, let $h_u(L)$ denote the number of messages sent during searching. Assuming that the advertising network $A(V_a, E_a, y)$ is a L-network, the corresponding number of messages will be equal to the number of the network links E_L. Eventually, the *total number of messages* $H(L)$ required for information dissemination (for both advertisement and searching), is given by,

$$H(L) = E_L + \sum_{\forall u \in J} h_u(L). \quad (1)$$

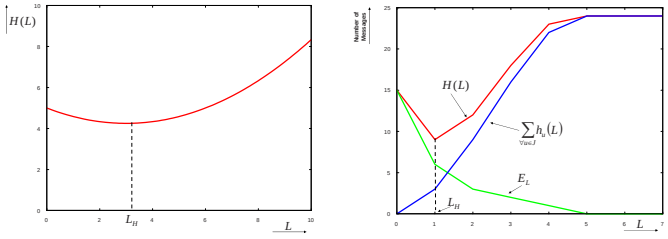
The study of the aforementioned expression of $H(L)$ would provide for the particular value of L, denoted as L_H, for which the minimum number of messages (i.e., $H(L_H)$) is assumed. Unfortunately, it is not possible – in the general case – to derive an analytical expression for E_L and $h_u(L)$ and therefore, it is not possible – in the general case – to derive a closed expression for L_H. On the other hand it is possible to make some interesting observations. It is important to note that E_L depends on the advertisement policy and $\sum_{\forall u \in J} h_u(L)$ on the searching policy. The complementary nature of both phase will be carefully investigated next by studying both elements of Equation (1).

As L increases, E_L decreases. For $L = 0$ (which is the case that all network nodes are aware about the information node s), $E_L = N + 1$. For large values of L (e.g., $L = N - 1$ for the special case of a line topology where the information node is located at the end point of the line), $E_L = 0$ (i.e., no need to advertise since the network nodes will look for it in the entire network). On the other hand, as L increases, $h_u(L)$ increases. For $L = 0$, $h_u(L) = 0$. For large values of L as before, $h_u(L) = |E|$. Eventually, $\sum_{\forall u \in J} h_u(L) = |J||E|$.

The complementary nature of both processes allows for different scenarios as L increases. The focus next is on the particular case case for which a *global minimum* of $H(L)$ is assumed for $L = L_H$, $L_H > 0$, as depicted in Figure 3.a. The derivation of L_H is based on the attempt to minimize the overall number of messages in the network $H(L)$ (constraint efficiency), given by Equation (1). L_H may not be suitable for those cases that searching has to terminate within a certain time period (i.e., constraint promptness). Suppose, for example, that searching is required to terminate within T time units. Assuming that the messages are sent during searching require one time unit to be processed and forwarded accordingly, and if the notification message about the information node location requires one time unit (sending back a message is faster since it does not require to be processed in the intermediate nodes), it is evident that it takes at most $L + 1$ time units to retrieve the information and terminate the searching. Consequently, $L + 1 \leq T$ should be satisfied in order for the promptness criterion to be satisfied.

Let $L_T = T + 1$ denote the upper bound of L for which the promptness criterion is satisfied. If $L_H < L_T$ and L is set to L_H, then the promptness criterion is satisfied and the total number of messages $H(L)$ is minimized. If $L_H > L_T$, then it is not possible to achieve minimum number of messages and at the same time satisfy the promptness criterion. For this particular case, the most suitable selection for L is L_T, in order to satisfy the promptness criterion and at the same time

the total number of messages $H(L)$ to be small as possible (even though not minimum since for this case $L_H \neq L_T$). Figure 3.a is helpful with respect to the aforementioned two cases.



a. b.

Fig. 3. Example of $H(L)$ as a function of L and number of messages for advertisement (E_L) and for searching ($\sum_{\forall u \in J} h_u(L)$).

Figure 3.b provides simulation results for a small grid network of 16 nodes for various values of L and for $J = 1$. The number of messages both for advertisement (E_L) and searching ($\sum_{\forall u \in J} h_u(L)$) are depicted. The results are in accordance with the previous analysis (as L increases, E_L decreases until 0 and $\sum_{\forall u \in J} h_u(L)$ increases until a certain value). $H(L)$ appears to assume a minimum value for $L = 1$.

V. Conclusions and Future Work

In this paper the information dissemination ingredients – advertisement and searching – have been investigated. Advertisement – responsible to disseminate certain pieces of information to a certain subset of the network nodes – is complementary to searching in the sense that the less intensive the advertisement, the more intensive the searching should be.

Therefore, in order to facilitate the design of an effective information dissemination mechanism, a set of criteria (i.e., correctness, promptness, fairness) were introduced. Eventually, given a searching mechanism and intensity, the (complementary) advertisement phase can be designed and parameterized so that the combined (complementary) phases meat the aforementioned criteria. Therefore, searching was considered as employing L-flooding – in the area of at most L hops away from any network node – and consequently, advertisement was responsible for disseminating information at most L hops away from any network node.

In the sequel, the conditions were studied under which the resulting dissemination information mechanism satisfies the aforementioned criteria. Many of the existing approaches (e.g., flooding-based, random walks) may be utilized for the aforementioned advertisement phase. When minimization of the number of messages sent during the advertisement phase is required, this turns out to be a dominating set problem that is known to be NP-hard and requires global network knowledge. Consequently, this is not a suitable approach for network environments with normally large number of nodes, due to the introduced high overhead both in time and network resources.

Future work will focus on further evaluation of the approach considered here using simulation results. Approaches like random walks, replicated random walks, random walks with jumps, will be used for both advertisement and searching in order to derive the conditions under which a particular information dissemination approach satisfies the aforementioned criteria.

Acknowledgment

This work has been supported in part by the project ANA (Autonomic Network Architecture) (IST-27489), the PENED 2003 program of the General Secretariat for Research and Technology (GSRT) co-financed by the European Social Funds (75%) and by national sources (25%) and the NoE CONTENT (IST-384239).

References

[1] A. Segall, “Distributed network protocols,” IEEE Trans. In- form. Theory, vol. IT-29, Jan. 1983.
[2] F. Banaei-Kashani and C. Shahabi, “Criticality-based Analysis and Design of Unstructured Peer-to-Peer Network as “Complex Systems,” in Proceedings of the Third International Symposium on Cluster Computing and the Grid, 2003, pp. 51 - 358.
[3] D. Tsoumakos and N. Roussopoulos, “Adaptive Probabilistic Search for Peer-to-Peer Networks,” 3rd IEEE International Conference on P2P Computing, 2003.
[4] K. Oikonomou, and I. Stavrakakis, “Performance Analysis of Probabilistic Flooding Using Random Graphs,” The First International IEEE WoWMoM Workshop on Autonomic and Opportunistic Communications (AOC 2007), Helsinki, Finland, 18 June, 2007.
[5] A.O. Stauffer, and V.C. Barbosa, “Probabilistic Heuristics for Disseminating Information in Networks,” Networking, IEEE/ACM Transactions on Volume 15, Issue 2, Page(s):425 – 435, April 2007.
[6] Gnutella RFC, http://rfc-gnutella.sourceforge.net/, 2002.
[7] C. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker, “Search and Replication in Unstructured Peer-to-Peer Networks,” ICS 2002, 2002.
[8] C. Gkantsidis, M. Mihail and A. Saberi, “Hybrid Search Schemes for Unstructured Peer-to-Peer Networks,” IEEE Infocom 2005, 2005.
[9] N. B. Chang and M. Liu, “Optimal Controlled Flooding Search in a Large Wireless Network,” Third International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt’05), 2005, pp. 229 - 237.
[10] V. Kalogeraki, D. Gunopoulos and D. Zeinalipour-Yazti, “A Local Search Mechanism for Peer-to-Peer Networks,” in CIKM (International Conference on Information and Knowledge Management), 2002.
[11] B. Williams, D. Mehta, T. Camp and W. Navidi, “Predictive Models to Rebroadcast in Mobile Ad Hoc Networks,” IEEE Transactions on Mobile Computing (TMC), 2004, Vol. 3, pp. 295-303.
[12] B. Williams and T. Camp, “Comparison of broadcasting techniques for mobile ad hoc networks,” ACM Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC), 2002, pp. 194-205.
[13] S, Boyd, A, Ghosh, B. Prabhakar and D, Shah, “Randomized gossip algorithms,” IEEE/ACM Trans. Netw., 2006, Vol. 14, pp. 2508-2530.
[14] A. Ganesh, L. Massoulie, D. Towsley, “The effect of network topology on the spread of epidemicss,” 13-17 March 2005, INFOCOM 2005, Vol. 2, 2005, pp. 1455-1466.
[15] Z. J. Haas, J. Y. Halpern and L. Li, “Gossip-based ad hoc routing,” IEEE/ACM Trans. Netw., 2006, Vol. 14, pp. 479-491.
[16] L. Tzevelekas, and I. Stavrakakis, “Improving Partial Cover of Random Walks in large-scale Wireless Sensor Networks,” 3rd IEEE WoWMoM AOC Workshop, 15 June 2009, Kos island, Greece, 2009.
[17] D. Kogias, K. Oikonomou, and I. Stavrakakis, “Study of Randomly Replicated Random Walks For Information Dissemination Over Various Network Topologies,” 6th WONS, February 2-4, 2009. Snowbird, Utah, USA.
[18] H.-Y. Yang, C.-H. Lin, and M.-J. Tsai, “Distributed Algorithm for Efficient Construction and Maintenance of Connected k-Hop Dominating Sets in Mobile Ad Hoc Networks,” IEEE Transactions on Mobile Computing, Vol. 7, No. 4, April 2008.