Wireless Pers Commun (2015) 84:1689–1722
DOI 10.1007/s11277-015-2600-2

Efficient Internet Access Framework for Mobile Ad Hoc Networks

Radwa Attia[1] · Rawya Rizk[1] · Hesham Arafat Ali[2]

Published online: 24 April 2015

Abstract The development of the Internet services and applications and the trend in the fourth generation (4G) wireless networks to all-IP networks have led to a growing demand for enabling Mobile Ad hoc NETworks (MANETs) to connect to the Internet and achieving the goal of omnipresence Internet which is accessible anytime and anywhere. However, such integration gives rise to a number of challenges. In this paper, an efficient framework which has considered six main challenges encountered in the MANET–Internet integration is proposed. An adaptive distributed multipath internet gateways discovery protocol is proposed, where the gateways advertise their information dynamically to an adaptive limited number of mobile nodes. For mobile nodes that cannot receive this information, an efficient dynamic routing algorithm for MANETs–Internet integration derived from the ant colony optimization algorithms is proposed. In addition, an improved Quality of Services based gateway selection mechanism providing load-balancing is proposed. The simulation study confirms that the proposed framework is scalable and able to cope with changes on the number and mobility of active sources connecting to the Internet and outperforms other conventional approaches in terms of end-to-end delay, packet delivery ratio while attaining acceptable overhead and fair load distribution among all gateways.

Keywords Ant colony optimization · Gateway discovery · Load balancing · MANET–Internet integration · QoS

✉ Radwa Attia
radwa_6007@yahoo.com

Rawya Rizk
r.rizk@eng.psu.edu.eg

Hesham Arafat Ali
h_arafat_ali@mans.edu.eg

[1] Electrical Engineering Department, Port Said University, Port Said, Egypt

[2] Computer Engineering and Systems Department, Mansoura University, Mansoura, Egypt

1 Introduction

The rapid evolution in the field of mobile computing is driving Mobile Ad hoc NETworks (MANETs) [1–3], in which Mobile Nodes (MNs) form a self-creating, self-organizing and self-administering wireless network. The development of the Internet services and applications and the trend in the fourth generation (4G) wireless networks to all-IP networks have led to a growing demand for enabling MANETs to connect to the Internet and achieving the goal of omnipresence Internet which is accessible anytime and anywhere [4, 5]. However, such integration gives rise to a number of challenges. This integration comprises several phases; a gateway discovery that includes Internet GateWay (IGW) solicitation and IGW advertisement, gateway selection, registration, forwarding, and handoff.

Recently, MANET–Internet integration has received wide interest from researchers [6–9]. In gateway solicitation, none of the proposals in the literature use route optimization algorithm. They update the traditional MANETs routing protocols such as AODV, DSR, etc. [6]. Since that the ant colony optimization (ACO) algorithms proved their superiority on the traditional MANETs routing protocols [10–12], a new GateWay SOLicitation (GWSOL) procedure inspired by the ACO algorithms is proposed. It outperforms the traditional routing protocols in terms of end-to-end delay, packet delivery ratio and routing overhead.

In gateway advertising, many protocols have been presented to adjust the proactive area [6], however they have many drawbacks. The main drawbacks are, some of them don't consider the present network conditions, others perform a central adjustment at the IGWs and/or consume very large overhead which in turn causes performance degradation. In this paper, a new protocol that adapts its behavior in a distributed manner, based on the present network conditions, and provides efficient and faster discovery for the gateways without consuming large overhead is proposed. In addition, there are no proposals for adjusting the time for the GateWay ADVertisement (GWADV) messages except some proposals that use fuzzy logic, which are very lack in terms of accuracy [6]. In this paper, an optimized feedback control system that enables the IGWs to adjust their GWADV interval dynamically according to the current network conditions is proposed.

In gateway selection, some proposals, select the IGW based on either the shortest number of hops or the IGW offered load [6], however, considering either of these metrics results in ignoring other important metrics and gives inaccurate selection process. In this paper, an improved Quality of Services (QoS) based IGW selection mechanism is proposed. It considers multiple metrics to select an optimum IGW. These metrics are, path quality in terms of bandwidth and queue length, gateway traffic load, and number of hops.

Each of the introduced proposals in the MANET–Internet integration considers only one of its phases without consideration of the complete solution for the integration. This was the main motive to find a complete solution that finds an answer for the following issues: (i) How to improve the packet delay and throughput that is affected by the IGW discovery time and handover delay. (ii) How to minimize the overhead of mobile IP and MANET routing algorithm between the Internet and MANETs. (iii) How to improve the IGW selection and achieve a load balance. (iv) How to guarantee the flexibility of the Internet connectivity.

In this paper, an efficient Adaptive Distributed Multipath IGW (ADMIGW) framework that considers the main challenges encountered in the MANET–Internet integration is proposed. The proposed framework comprises:

(a) A new efficient and adaptive multipath gateway discovery protocol for MANETs–Internet integration inspired by the ACO algorithms.
(b) A new proactive area adjustment protocol, which combines the advantages of the three conventional IGW discovery protocols; proactive, reactive and hybrid protocols, and can provide efficient and faster discovery for the gateways.
(c) A new optimized feedback control system that adjusts the gateway advertisement time interval.
(d) In addition, an improved QoS-based IGW selection mechanism which reduces the data drop rate and provides load-balancing across a set of the discovered access gateways is proposed. It considers path quality, IGWs load, and hop count metrics to select an optimum IGW.

The remainder of the paper is organized as follows: Sect. 2 introduces the main challenges and related work for MANET–Internet integration. Section 3 presents the phases of the proposed framework for Mobile Ad Hoc wireless Internet access networks and the associated proposed protocol for each phase. An analytical evaluation is presented in Sect. 4. Simulation results are shown in Sect. 5. Finally, the conclusion is drawn out in Sect. 6.

2 Main Challenges and Related Work for MANET–Internet Integration

Generally, connecting MANETs to the Internet does not come without difficulties [6–9]. Six main important challenges of MANET–Internet integration are encountered in this paper.

2.1 MANET–Internet Routing

Communication between the Internet nodes and the MNs is done throughout specialized IGWs, which act as bridges between the MANET and the Internet. Some of the proposed routing solutions in the literature are presented in [13–16].

2.2 Gateway Discovery

Three main approaches have been proposed for the IGW discovery function [6, 17]: proactive, reactive, and hybrid approaches. In proactive approach, IGWs periodically broadcast GWADV messages throughout the MANET. For reactive approach, Active Sources (ASs) discover the IGWs itself by broadcasting a GWSOL message. In response to the solicitation, each IGW sends a special route reply message back to the MNs offering its services. In the hybrid approach, for MNs in a certain range around the IGW, proactive approach is used while MNs residing outside this range use reactive approach. Many adaptive approaches are introduced to improve the performance of the hybrid approach [6, 18–26].

2.3 Gateway Selection

Selection can take place either at the IGWs or on MNs. In the proxy approach, IGWs can selectively reply to route requests depending on a specific load and security policies. When the MNs make the selection themselves, a straightforward solution is to select the IGW that has the shortest number of hops. However, other metrics, like the IGW offered load can be used [6, 26–29]. Also MNs may choose a set of IGWs for multi-homing or load-balancing

[6, 30, 31]. In this case, the mechanism must also support forwarding to multiple gateways simultaneously.

2.4 Global Addressing

In order to be able to communicate with the Internet, an MN needs an address auto-configuration mechanism in order to configure a global routable and topological correct address. IPv6 defines two fundamental principles for auto-configuration [6, 32–37]; stateful and stateless auto-configuration. In stateful auto-configuration, IGW automatically assigns addresses to the requesting MNs and manages the address space. However, centralized approaches are not suitable for MANETs due to possible network partitions. In stateless auto-configuration, the IGW can advertise within its control messages a network prefix from which the MNs can derive an IP address. With stateless auto-configuration, there is a risk of setting duplicate addresses in a network, a Duplicate Address Detection (DAD) mechanism [6, 36, 38] may be used to solve this problem. Afterward, the global address of the MN should register with the selected IGW.

2.5 Forwarding

The IGW forwarding mechanism in the MANET plays a crucial rule for the flexibility of the Internet connectivity. It can be classified according to the tunneling mechanism used [6, 9, 31, 39]; tunneling and non-tunneling based integrated routing solutions.

2.6 Handover

The handover decision depends on the movement detection method. There are two methods for movement detection [6, 7, 40, 41], invalidating the route entry and receiving the GWADV messages. The MNs would be forced to register to an alternative IGW when the original IGW fails. Also, when an MN finds another best IGW, it initiates the handover to the new IGW for optimizing the route.

Considering the above challenges, lots of solutions have been proposed in literature. However, most of these solutions have a number of drawbacks. In this paper, a complete solution that considers all of these challenges in a hierarchal way is proposed.

3 Proposed Framework for Mobile Ad Hoc Wireless Internet Access Networks

This section presents the proposed framework that considers the main challenges encountered in MANET–Internet integration. The structure of the proposed framework combines six phases as depicted in Fig. 1.

3.1 Network Model

3.1.1 Network Architecture

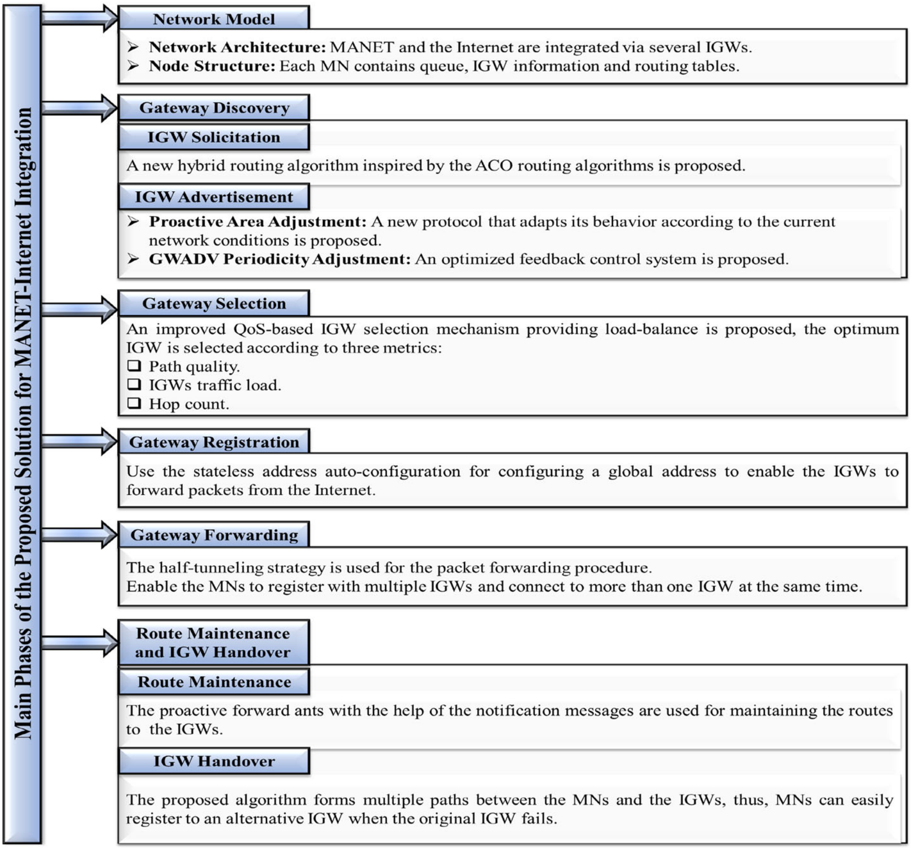
In the case of single IGW scenario, simultaneous utilization of the IGW by several MNs leads to heavy traffic congestion around the IGW. So in this work, a scenario where a

MANET is connected to the Internet via several IGWs is considered. IGWs are located between MANETs and the Internet and are capable of providing bidirectional global connectivity to the MNs which are connected to them either directly or via one or more intermediate nodes.

3.1.2 Node Structure

Every MN of type store-and-forward holds a queue, assuming that the links between any two MNs are bi-directional links. Furthermore, at each MN two tables are used to maintain the IGW information (the gateway information table and the routing table). Each entry in the gateway information table contains the IP address of the IGW, the lifetime of this entry and any other general information concerning the IGW. Whenever an MN receives a GWADV, it adds or updates the gateway information table and the routing table.

As indicated in Fig. 2, the rows represent the IGWs, and the columns represent the neighbor MNs that can be reached from that MN. The entries are numeric values between 0 and 1 which signify the pheromone values $\tau_{Gateway,Neighbor}$. The pheromone value τ_{in} which


Main Phases of the Proposed Solution for MANET-Internet Integration
Network Model

- **Network Architecture:** MANET and the Internet are integrated via several IGWs.
- **Node Structure:** Each MN contains queue, IGW information and routing tables.

Gateway Discovery
IGW Solicitation

A new hybrid routing algorithm inspired by the ACO routing algorithms is proposed.

- **Proactive Area Adjustment:** A new protocol that adapts its behavior according to the current network conditions is proposed.
- **GWADV Periodicity Adjustment:** An optimized feedback control system is proposed.

An improved QoS-based IGW selection mechanism providing load-balance is proposed, the optimum IGW is selected according to three metrics:
❑ Path quality.
❑ IGWs traffic load.
❑ Hop count.

Use the stateless address auto-configuration for configuring a global address to enable the IGWs to forward packets from the Internet.

The half-tunneling strategy is used for the packet forwarding procedure.
Enable the MNs to register with multiple IGWs and connect to more than one IGW at the same time.

The proactive forward ants with the help of the notification messages are used for maintaining the routes to the IGWs.

The proposed algorithm forms multiple paths between the MNs and the IGWs, thus, MNs can easily register to an alternative IGW when the original IGW fails.

Fig. 1 Main phases of the proposed framework for MANET–Internet integration

represents the probability of selecting n as the next MN when the IGW node is *i*, is saved with the constraint:

$$\sum_{n\ \in N_K} \tau_{in} = 1, \quad N_k = \{Neighbors(k)\} \tag{1}$$

3.2 The Proposed Multiple Gateway Discovery Protocol

The proposed multiple gateway discovery protocol comprises two main techniques; IGW solicitation and IGW advertisement.

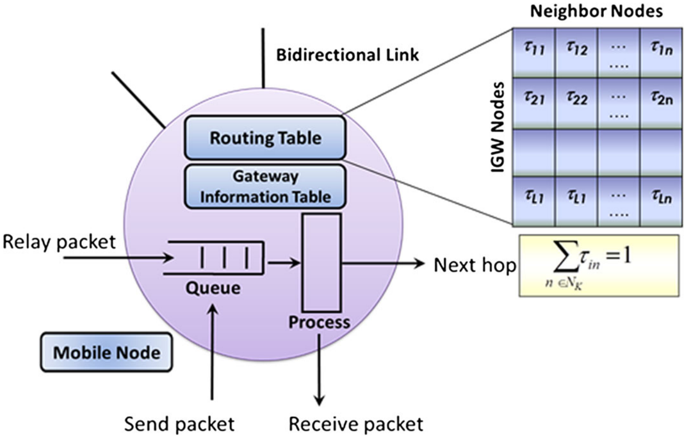


Fig. 2 Mobile node structure for MANET–Internet integration

Table 1 Notations used in the GWSOL algorithm

Notation	Meaning
$D_{i,j}$	Link propagation delay between two MNs *i* and *j*
$Q_{i,j}$	Queue length between *i* and *j* (number of bits waiting in the queue)
$\rho_{i,j}$	Utilization of the link between *i* and *j*
λ_j	Arrival rate at node *j*
μ_j	Service rate at node *j*
S_a	Size (bits) of the ant packet
$B_{i,j}$	Bandwidth (bps) of *i* and *j*'s link
$P_{i,j,G}$	Probability of MN *i* to choose *j* as the next hop on the way to the IGW *G*
$\tau_{i,j,G}$, $\eta_{i,j}$ and PL_j	Pheromone value, heuristic function and power level; respectively
H_i	Represents the neighbors of node *i*
$(\alpha,\ \beta,\ \gamma \geq 1)$	Weight functions that control τ, η, and *PL*; respectively
T_m *and* T_c	Minimum and current trip time of the RFA from an MN to the IGW

3.2.1 The Proposed Multiple Gateway Solicitation Algorithm

In gateway solicitation, all proposals present in the literature update one of the traditional MANETs routing protocols such as AODV, DSR, etc. [6]. None of them use route optimization algorithm. Since that the ACO algorithms proved their superiority on the traditional MANETs routing protocols [10–12], a new GWSOL procedure inspired by the ACO is proposed. It outperforms traditional routing protocols in terms of end-to-end delay, packet delivery ratio and routing overhead.

In the proposed GWSOL algorithm, whenever an MN wants to communicate with a fixed, wired Internet host, at first it has to find an IGW by searching its gateway information and routing tables for a fresh route to any IGW. If a route is found, the communication can be established; otherwise, the MN starts a multiple GWSOL procedure. It consists of two main steps, the reactive path setup step and proactive route maintenance step. Various notations used in this algorithm are shown in Table 1.

3.2.1.1 Reactive Path Setup Step This step consists of two main stages which are the reactive forward and backward agent stages.

Reactive Forward Agent Stage An AS node starts a communication session with the IGWs by generating a control packet called Reactive Forward Agent (RFA) in order to investigate paths to the several IGWs. In the proposed algorithm, the RFA is broadcasted only to neighbors one hops away from the AS node (N_{SN}). Several agents can be produced from one broadcast, those agents called “Same generation agents”, which have the same source address and generation number. For each MN, upon the reception of the RFA, they unicast it to one of their neighbors until it reached the IGWs. While traveling toward the IGWs, the RFAs keep memory of their paths and of the step-by-step time elapsed $T_{Current,Next}$ since the launching time. The required time for travelling of the packet from an MN i to a neighboring j can be calculated using the following formula:

$$T_{i,j} = D_{i,j} + (Q_{i,j} + S_a)/B_{i,j} \tag{2}$$

$$Q_{i,j} = \frac{\rho_{i,j}^2}{1 - \rho_{i,j}} \tag{3}$$

$$\rho_{i,j} = \frac{\lambda_j \times \tau_{i,j,d}}{\mu_j} \tag{4}$$

At every MN i, each RFA does the following:

i. If the MN i is the IGW node or has fresh routing information to any IGW, it will generate a Reactive Backward Ant (RBA).

ii. Else, the RFA checks whether node i address is in the stack. If present, there will be a loop.

  - If the loop lasted higher than half the RFA age (F_A), the RFA is discarded to avoid loops.
  - Else, the MNs in the loop are dropped from the RFA’s stack memory.
    Otherwise,
  - While the node i has a fresh routing information for one or more IGW, it decides the next hop j for the best IGW with a probability $P_{i,j,G}$.

Otherwise,

$$P_{i,j,G} = \frac{[\tau_{i,j,G}]^{\alpha}[\eta_{i,j}]^{\beta}[PL_j]^{\gamma}}{\sum_{e \in H_i}[\tau_{i,e,G}]^{\alpha}[\eta_{i,e}]^{\beta}[PL_e]^{\gamma}} \tag{5}$$

where,

$$\eta_{i,j} = 1 - \frac{Q_{i,j}}{\sum_{e \in H_i} Q_{i,e}} \tag{6}$$

- Else if node *i* does not have any routing information for any IGW, it selects one of its neighbor nodes with a probability $P_{i,j}$ that depends on their power level and queue occupancy.

$$P_{i,j} = \frac{[\eta_{i,j}]^{\beta}[PL_j]^{\gamma}}{\sum_{e \in H_i}[\eta_{i,e}]^{\beta}[PL_e]^{\gamma}} \tag{7}$$

When RFAs reached any IGW, the IGW will take the decision whether to accept the request or not according to its own load. The IGW can accept a limited number of same generation agents (A_g) and generate RBAs that inherits the memory of the RFAs for supporting a multi-path between the AS and the IGWs.

Reactive Backward Agent Stage The RBA retraces the path of its related RFA. If this is not possible because the next hop is not there, the RBA is discarded. This agent adds or updates the gateway information and routing tables of MNs on this path based on the RFA trip time. To ensure that the sum of the pheromone values in each row remains 1, the routing table of each intermediate MN is updated according to the following rules:

If the node *i* was in the path of the RFA then, the pheromone value will be increased,

$$\tau_{new}(i) = \tau_{old}(i) + r \times [1 - \tau_{old}(i)] \tag{8}$$

Else, it will be decreased,

$$\tau_{new}(i) = \tau_{old}(i) - r \times \tau_{old}(i) \tag{9}$$

where $\boldsymbol{r = T_m/T_c}$ (between 0 and 1) is the reinforcement factor which expresses the path quality.

The RBAs are produced by the IGWs. However, an intermediate MN can generate a reply packet and sets the flag for the RBA generation as a response to a request for the IGW information only if it has a valid IGW entry in the gateway information table and a valid route cache entry for that IGW, where the term ‘valid’ means “not expired and the information or path can be used without validation”. When all RBAs reach the AS node there will be multiple paths to different IGWs. The details of the reactive path setup phase are illustrated in Algorithm 1. The AS node will select one of these IGWs and begin the registration process to the selected IGW.

3.2.1.2 Proactive Path Maintenance Phase Congestion in a network may occur at any interval, when the battery power of the MNs decreased and also when the number of packets coming in an MN exceeds its buffer capacity. Thus, during data transmission, an AS periodically produces Proactive Forward Agents (PFAs) at the rate depends on the data sending rate to maintain and enhance the established paths and to discover better or alternative paths. The PFA performs the same role as the RFA, it gathers fresh information about the established path and updates the pheromone values in the MNs routing tables by the corresponding Proactive Backward Agents (PBAs).

Additionally, if an MN detects a link failure by the unsuccessful transmission of the data packets, and there is no alternate path for this packet, then the MN dispatches a forward Path Repair Agent (PRA). The forward PRA makes the same function as the RFA, any MN receives this packet and have the latest information about the IGW can generate a backward PRA for the requested MN.

Algorithm 1 Reactive Path Setup Algorithm

Input: IGW request for an AS *S* (GWSOL(*S,G*))

Output: Multiple paths from *S* to different IGWs

```
Procedure GWSOL(S,G)
  S broadcast copies of RFAs to all $N_{SN}$
  for each neighbor ∊ $N_{SN}$
      i=1
      while (Current_MN ≠ G or have fresh information for G)
          if loop detected
              if $loop>0.5 \times F_A$
                  Destroy_ RFA
              else
                  Drop all MNs in the loop and destroy all their stack memory
              end if
          else
              Next_hop_MN := Select one neighbor
              List_crossed_MNs (i) := Current_MN
              i++
              Current_MN : = Next_hop_MN
          end if
      end while
      Generate RBA(List_crossed_MNs)
      Kill_RFA
      While (Current_MN ≠ S)
          Update routing and gateway information tables(Current_MN,G)
          Next_MN := List_crossed_MNs(i)
          i=i-1
          Current_MN := Next_MN
      end while
  end for
  Kill_RBA
end Procedure
Begin Procedure GWSEL(S,G)
```

3.2.2 The Proposed Gateway Advertising Protocol

The main important factors involved in sending unsolicited GWADVs are the adjustment of the proactive area, defined by the Time-To-Live (TTL) field value of the GWADV, and the Time (T) interval between sending two consecutive advertisements, based on the network traffic conditions.

3.2.2.1 Adjustment of the Proactive Area Many protocols have been presented in the literature to adjust the proactive area [6, 21–24], however they have many disadvantages. Some of them don't consider the present network conditions which is very crucial. Others perform a central adjustment at the IGWs and/or consume very large overhead which in turn causes performance degradation. In this paper, a new protocol that adapts its behavior in a distributed manner, based on the present network conditions, and can provide efficient and faster discovery for the gateways without consuming large overhead is proposed.

Initially, any MN receives or generates an RFA or a PFA message or a PRA, or a Route ERRor (RERR) message, or it already relaying data packets, establish itself as a Needy MN (NMN) i.e., an MN in the need of the GWADVs, and updates the timer used for the NMN indicator.

For each IGW, during the periodic or the adaptive GWADVs, at first the IGW broadcasts the GWADV to neighbors (N_{GN}) only one hop away from it (with TTL = 1). After that, on the reception of the GWADV message, each node verifies whether its neighbors are NMNs or not:

- If at least one is a NMN, it forwards the GWADV further only to the NMNs, to ensure that the GWADVs are propagated only to nodes already in the need of the GWADVs.
- Else, it unicasts the GWADV to the MN having the maximum number of neighbors, to cover almost all MNs in the network.

Each MN has a permission to forward a limited number of the GWADVs. The GWADV packet carries a counter that is incremented by one whenever an MN is visited. The GWADV packet is dropped when the counter exceeds the certain threshold value equal to the network diameter (D_N). Algorithm 2 shows the main processes of the proactive area adjustment.

Algorithm 2 Proactive Area Adjustment

Input: IGW information

Output: MNs get the latest IGWs information

```
Procedure GWADV(G,MNs)
  IGW broadcast copies of GWADVs to all $N_{GN}$
  TTL=1
  While (TTL<=$D_N$)
      for each MN receives GWADV
          Current_MN checks its neighbors
              if (NMN==NULL)
                  Unicast the GWADV to the MN having the maximum number of neighbors
                  Update routing and gateway information tables
              else
                  Multicast GWADVs to all NMNs
                  Update routing and gateway information tables
              end if
          TTL++
      end for
  end while
end Procedure
```

The proposed protocol is further explained with the help of graph as shown in Fig. 3. As shown, all ASs receives the GWADV message, traversing hop-by-hop through all the NMNs. The reason of unicasting the GWADV although the absence of the NMN is as shown, there may be some NMN or ASs far from the IGW or may any non-NMN becomes a NMN or an AS after a little time. This characteristic of the proposed protocol helps the MNs to quickly learn the route towards the IGW in a dynamic network, without generating an excessive overhead.

3.2.2.2 Adjustment of GWADV Interval In the literature, there is almost no proposal for adjusting the time for the GWADVs messages except some proposals that use fuzzy logic, which are very lack (principally in terms of accuracy) [6]. Thus, in order to improve the overall network performance, especially in terms of overhead, an optimized feedback control system is proposed. It should be installed in the IGWs to enable them adjust their GWADV interval dynamically according to the current network conditions. Two main phases are considered in this system, IGWs information collections and the feedback control phases.

- IGWs Information Collections Phase

IGWs periodically gather some specific information metrics that capture the current state of the network these metrics are:

(1) Link Connectivity

Link Connectivity (L_C) [42] is the mobility metric that demonstrates the link connectivity duration two hops away from the IGWs. This metric is in inverse proportion to the relative velocity between the IGWs/MNs and their neighbor MNs and related to the number of these neighbors.

If N_{GN} is the number of MNs one hops away from the IGW, average link connectivity L_C of IGW G over a time t_c seconds is defined by [42]:

$$L_C = \int_{0}^{t_c} \frac{1}{N_{GN}} \sum_{I}^{N_{GN}} l_c(G, I, t)dt \tag{10}$$

where $l_c(G,I,t)$ is the link connectivity duration between the IGW G and a specific neighbor node $i \in N_{GN}$ at time t and can be calculated as follows:

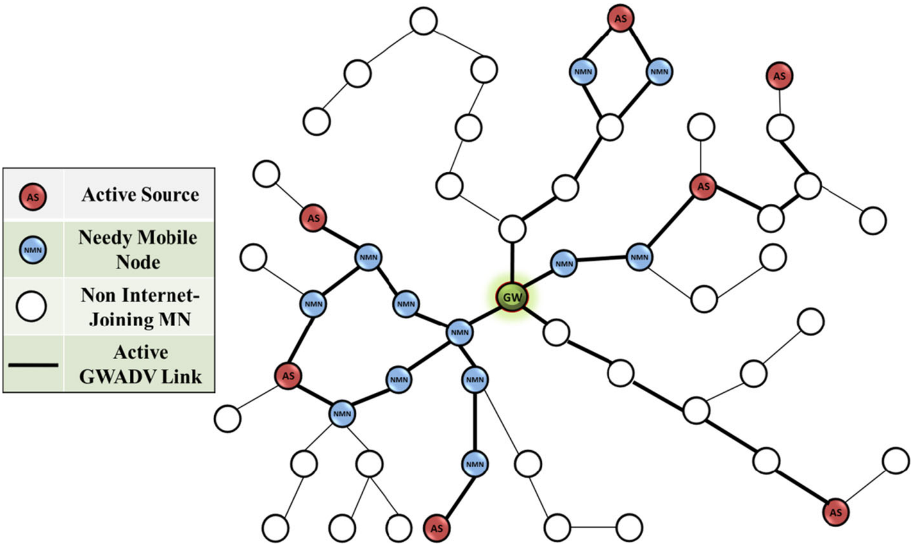


Fig. 3 The proposed adaptive gateway advertisement protocol

$$l_c(G, I, t) = e^{-av(G,I,t)} \tag{11}$$

where *v(G,I,t)* is the relative velocity between G and I at time t:

$$v(G, I, t) = \sqrt{b^2 + c^2} \tag{12}$$

$$b = v(G, t)\cos\theta(G, t) - v(I, t)\cos\theta(I, t) \tag{13}$$

$$c = v(G, t)\sin\theta(G, t) - v(I, t)\sin\theta(I, t) \tag{14}$$

where *v(I,t)* and *Θ(I,t)* are the speed and direction of node *I* at time *t*; respectively.

The IGW neighbors also calculate the LC value, as in Eq. 10, and sent it periodically to their related IGWs to enable each IGW estimate its two hops away paths strength. The IGW calculates a normalized value (L_{CN}) for the link duration using the following formula:

$$L_{CN} = \frac{Avg(L_C) - Min(L_C)}{Max(L_C) - Min(L_C)} \tag{15}$$

(2) Traffic Loss

The IGWs also monitor the amount of the incoming traffic loss during a specified time interval. In this paper, traffic loss metric mainly reflects the traffic load, mobility and congestion status of distant MNs because the ants have already selected good quality paths in terms of energy and queue length, which are the main items result in packet loss. Traffic loss (T_L) for a path *P*(*S*,*G*) between an AS *S* and an IGW *G* can be represented mathematically by indirect multiplicative metrics:

$$T_L = 1 - \prod_{N \in P(S,G)} (1 - L(N)) \tag{16}$$

where *L(N)* is the traffic loss at MN N. The IGW capture the traffic loss for all paths P_N and then measure the amount of all incoming traffic losses according to:

$$T_{LN} = \frac{\sum_{P=1}^{P_N} 1 - \prod_{N \in P(S,G)} (1 - L(N))}{P_N} \tag{17}$$

(3) Routing Error

As discussed earlier, the PRA messages are dispatched by MNs that detects link failure during the transmission of the data packets. IF the IGW receives these messages, it concludes that there is a high mobility in the whole network because there is no other intermediate MN has the latest information about it. In this paper, if N_{PRA} and N_{GWSOL} are the number of received PRAs and GWSOL messages; respectively, during a specified time interval, an error routing (E_R) metric can be computed as follows:

$$E_R = \frac{N_{PRA}}{N_{PRA} + N_{GWSOL}} \tag{18}$$

(4) MNs Traffic Load

The IGWs should capture the number of active links (N_{AL}) relaying data packets for ASs. This metric reflects the number of MNs loaded with data packets, with the increase of this number, there is a need for the IGWs information to avoid losses. The MNs traffic load (MNL) metric can be computed as follows:

$$MN_L = \frac{N_{AL}}{N_{AL} + N_S} \tag{19}$$

- Feedback Control Phase

Feedback control theory [43] is used to control dynamic and unpredictable systems. Thus, an optimized feedback control system is installed in the IGWs to enable them periodically and dynamically adjust their GWADV interval according to gathered metrics in the previous phase and the past temporal network status information. The proposed feedback control system is shown in Fig. 4.

As shown in Fig. 4, the Calculator takes the several input metrics and calculates the amount of changes in the network status (ΔI_{New}) as follows:

$$\Delta I_{New} = a_1 \times L_{CN} + a_2 \times T_{LN} + a_3 \times E_R + a_4 \times MN_L \tag{20}$$

where *a1*, *a2*, *a3* and *a4* are the weighting factors for L_{CN}, T_{LN}, E_R and M_{NL}; respectively,

$$0 \le a_1,\ a_2,\ a_3,\ a_4 \le 1 \quad and \ \sum_{i=1}^{4} a_i = 1.$$

The Comparator computes the difference between the output of the calculator (ΔI_{New}) and the value of ΔI in its previous cycle (ΔI_{Old}), to estimate how much the network status change are,

$$\Delta I = \Delta I_{Old} - \Delta I_{New} \tag{21}$$

This value is then feedback to the function block ΔI_{Old} to replace the ΔI_{Old} value with ΔI_{New} for the next estimation interval as follows:

$$\Delta I_{Old} \leftarrow \Delta I_{New} = \Delta I_{Old} - \Delta I \tag{22}$$

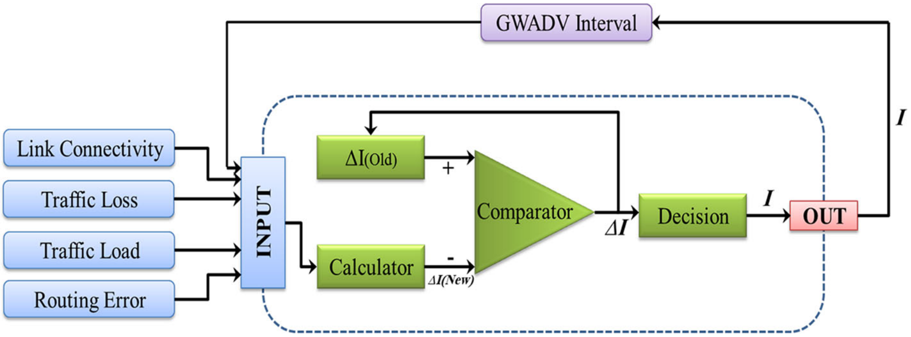


Fig. 4 The proposed feedback control system

Also, this value is used as input to the Decision function block to decide the next GWADV time interval according to the following rules:

$$I = \begin{cases} I_{Max} & \Delta I > Th_{Max} \\ I_{Old} + \Delta I \times I_{Old} & Th_{Max} > \Delta I > 0 \\ I_{Old} - |\Delta I| \times I_{Old} & Th_{Min} < \Delta I < 0 \\ \quad I_{Min} & \Delta I < Th_{Min} \end{cases} \tag{23}$$

As shown, when the amount of change ΔI is greater than a specific positive maximum threshold value (Th_{Max}), it means that the network is currently stable and there is no need for the fast transmission of the GWADV, thus the next time interval is set to the maximum (I_{Max}). On the other hand, when ΔI is lower than a specified negative minimum threshold value (Th_{Min}), this means that the network is highly unstable and there is a critical need for the GWADV, thus the time interval is set to the minimum (I_{Min}). Otherwise, the time interval is increased or decreased according to the amount of change, as shown in rule 23. The main processes of the system are shown in Algorithm 3.

Algorithm 3 GWADV Interval Adjustment

Input: L_{CN}, T_{LN}, E_R and MN_L information metrics, I_{Old}, ΔI_{Old}, Th_{Max} and Th_{Min}

Output: Decide the next GWADV time interval I_{New}

```
Procedure Interval (G)
    Calculate $\Delta I_{New} = a_1 \times L_{CN} + a_2 \times T_{LN} + a_3 \times E_R + a_4 \times MN_L$
    $\Delta I = \Delta I_{Old} - \Delta I_{New}$
    if $\Delta I > Th_{Max}$
        $I_{New} = I_{Max}$
    else if $\Delta I < Th_{Min}$
        $I_{New} = I_{Min}$
    else if $0 < \Delta I < Th_{Max}$
        $I_{New} = I_{Old} + (\Delta I \times I_{Old})$
    else
        $I_{New} = I_{Old} - (\Delta I \times I_{Old})$
    end if
end procedure
Begin Procedure GWADV(G,MNs) upon the end of $I_{New}$
```

This proposed system effectively manages the GWADV interval according to the network conditions in order to not cover the network with GWADV messages needlessly and inject a high amount of overhead. And on the other hand, not store stale routing information within the MNs, while still providing the MNs with a good chance of finding a better path to the previously used IGW or to an even better IGW and also to permit MNs to maintain up-to-date IGWs information.

3.3 The Proposed Gateway Selection Mechanism

Gateway selection can take place either at the IGWs or on MNs. When the MNs make the selection themselves, a straightforward solution is to select the IGW that has the shortest number of hops [6], which is a very poor metric. Other metrics like the IGW offered load are used in [26–29], however, the path quality metric is not included and/or they do not provide a complete solution for the selection mechanism.

In this section, an improved QoS-based IGW selection mechanism is proposed. It considers multiple metrics to select an optimum IGW, taking into account that this selection will be applied on the multipath established by the agents which are already selected according to the pheromone values, the queue length and power level, i.e. good quality paths. This mechanism selects the optimum IGW according to:

i. Trip Time (TT): Used to regulate which IGW has a good quality path (in terms of bandwidth and queue length).
ii. Gateway traffic Load (GL): Used to avoid choosing the overloaded IGWs, decrease the processing latency of packets from/to the Internet and balance traffic load among all IGWs.
iii. Hop Count (HC): Used for fast convergence as well as thriftiness of resources. Also, a packet routing through a shorter path will have a better chance to face less collisions and congestion.

To combine these three metrics as one comparable metric, one of the Multiple Attribute Decision Making (MADM) techniques [44] called Simple Additive Weighting (SAW) method is used to outrank the optimum IGW by computing the weighted sum of all metric values. The SAW method specifies how attribute information is to be processed in order to arrive at a choice. The decision matrix A in the SAW method has four main parts, namely: (a) alternatives or gateways, G_i *(*for $i = 1, 2, \ldots., n$*)* (b) attributes or metrics, B_j *(*for $j = 1, 2, 3$*)* (c) weight or relative importance of each attribute, w_j and (d) value of each metric, M_{ij}.

$$A = \begin{array}{c} \\ G_1 \\ G_2 \\ . \\ . \\ G_n \end{array} \begin{array}{c} \begin{array}{ccc} B_1 & B_2 & B_3 \end{array} \\ \begin{bmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & M_{23} \\ . & . & . \\ . & . & . \\ M_{n1} & M_{n2} & M_{n3} \end{bmatrix} \end{array} \tag{24}$$

The task of the decision maker here is to rank the entire set of IGWs and find the best IGW. There are three fundamental steps of the SAW method: scaling the comparable value, applying the weighting factors, and computing the overall performance score of an alternative.

3.3.1 Scaling the Comparable Value

Firstly, all the elements in the decision matrix must be normalized to the same units, so that all possible metrics in the decision problem can be considered. Metrics are scaled either positively or negatively according to Eqs. 25 and 26; respectively. In the proposed gateway selection method all metrics are scaled negatively because it is better to have those values lower. When the target measures are unified, a new evaluation matrix B (27) is getting.

$$a_{ij} = \frac{M_{ij} - \min\{\mathrm{M_{ij}}\}}{\max\{M_{ij}\} - \min\{M_{ij}\}}, \quad a_{ij} \in [0, 1] \tag{25}$$

$$a_{ij} = \frac{\max\{\mathrm{M_{ij}}\} - \mathrm{M_{ij}}}{\max\{M_{ij}\} - \min\{M_{ij}\}}, \quad a_{ij} \in [0, 1] \tag{26}$$

$$B = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ . & . & . \\ . & . & . \\ a_{n1} & a_{n2} & a_{n3} \end{bmatrix} \tag{27}$$

3.3.2 Applying the Weighting Factors

The original AS node can simply select a weighting factor w_1, w_2 and w_3 according to the priority and relative importance of TT, GL and HC; respectively where:

$$0 \le w_1, w_2, w_3 \le 1 \quad \mathit{and} \ \sum_{j=1}^{3} w_j = 1 \quad \mathit{and} \ w = [\, w_1 \quad w_2 \quad w_3 \,]$$

3.3.3 Computing the Overall Performance Score

The last step is to calculate the final weight matrix (W) for all IGWs candidates, which combines the three metrics by multiplying the evaluation matrix and the weighting factors, as shown in (28).

$$W = \begin{bmatrix} W_1 \\ W_2 \\ : \\ W_n \end{bmatrix} = B \times w^T \tag{28}$$

Finally, after obtaining the W_i value of each candidate IGW, the IGW with the highest value will be selected and the registration process will begin with it. Algorithm 4 shows the main IGW selection steps.

Algorithm 4 IGW Selection Algorithm

```
Input: Multiple paths $P_j$ with normalized TT, GL and HC information for each $P_j$
Output: Select the optimum $G$
Procedure Select($S, G$)
    Calculate $W_1 = 0.5 \times TT_1 + 0.3 \times GL_1 + 0.2 \times HC_1$ for the first path $P_1$
    for $i = 2 : j$
        $W_i = 0.5 \times TT_i + 0.3 \times GL_i + 0.2 \times HC_i$
        if $W_i > W_1$
            $W_1 = W_i$
        end if
    end for
    Select G and P related to $W_1$
end Procedure
Begin registration to G along P
```

3.4 Gateway Registration

After the selection of the IGW, the AS node uses a stateless address auto-configuration. It uses the selected IGW's prefix and host address part of local address to auto-configure a global address. Because the local address is certainly unique at the phase of MANET-local address auto-configuration, the global address is also expected unique within the MANET.

The global address of the AS should register with the selected IGW to enable the IGW to appropriately forward packets from/to the Internet. The AS sends a registration request message to the IGW; this message is forwarded based on the selected path and the pheromone values in the routing tables. Then, the IGW returns a registration acknowledgment message to it and add an entry in the MANET global address table. The MANET global address table also enables the IGW to find out duplication of the global addresses and in turn, The IGW can give assistance to detect the duplication of the local address within the MANET. In case of the failure of the registration for a number of times, the AS node removes the respective IGW table entry and select a new IGW.

3.5 Gateway Forwarding

The half-tunneling strategy [9] is used in this paper, where the traffic from the MANET domain to the Internet uses tunneling or encapsulation [31], while traffic from the Internet to the MANET domain uses MANET forwarding without tunneling, see Fig. 5.

As shown in Fig. 5, with half-tunneling, an encapsulated packet from an AS node A to an Internet destination B is sent to the selected IGW according to the pheromone values in the routing tables. At the IGW, the packet exits the tunnel and is decapsulated before forwarding it to B. Return traffic from B does not need to be tunneled, since the AS IP address is routable within the MANET and also to reduce the overhead of adding additional IP header for the tunneling.

The main advantage of the half-tunneling solution is that the AS node of a flow is always in complete control and alone carries the entire state essential to forward the packets to the IGWs. Also, the tunneling process efficiently can make use of multiple IGWs at once for the benefit of multi-homing/load balancing or performing soft handovers, see Fig. 6.

3.6 Route Maintenance and Gateway Handover

3.6.1 Route Maintenance

For route maintenance, to guide the RFAs and the PFAs better, each MN periodically informs its neighbors of its existence by broadcasting a NOtification Message (NOM), short messages containing only the sender address, to them every time t_{NOM}. If an MN receives a hello from a new MN, it adds it in its routing table. After missing a certain number of notifications (allowed-notification-loss = 2), it is removed. Thus, these messages allow MNs to clean up stale entries from their routing tables when detecting broken links.

3.6.2 Gateway Handover

The ASs would register to an alternative IGW when the original IGW fails or they find another best IGW. However, the IGW handover procedure due to the existence of a better IGW result in a significant delay and may interrupt the ongoing sessions. Thus, the half-tunneling procedure provides the MNs to register with multiple IGWs and connect to more than one IGW at the same time. When an MN creates a new session, the MN can choose an

IGW which is different from that of the old session. In other word, each session independently chooses its connecting IGW. Once an IGW is chosen for the session, the ongoing session will not change the IGW unless the IGW fails.

As discussed earlier, the MN dispatches the PRA if it detects a link failure by the failed transmission of data packets. In this case, if no backward PRA is received within a certain time period, then the MN concludes that the path repair failed. Then, it discards all the temporarily buffered packets and sends a link failure notification (RERR) message to the AS about the lost IGW. The AS removes the respective IGW table entry and select a new IGW, if there exist, otherwise, it begins a new IGW solicitation procedure.

4 Analytical Evaluation of AMPIGW and Other Gateway Discovery Approaches

In this section, the analytical model for computing the IGW discovery overhead for the three IGW discovery approaches and for the proposed ADMIGW protocol is provided. The IGW discovery overhead is a performance metric that measures the scalability of the IGW discovery approaches and can be computed as the total number of control messages associated with the IGW discovery procedure. The analysis assumes that all hosts generate the same traffic pattern per time interval. Table 2 depicts the basic parameters which had been used in the model.

4.1 Proactive IGW Discovery Overhead

In the proactive IGW discovery approach, IGWs periodically broadcast their GWADV messages to the whole MANET regardless the number of ASs. In this paper, the total IGW discovery overhead Θ_P in the number of messages is computed as follows:

$$\Theta_P = N_{adhoc} \times N_G \times (\lambda_{GA} \times t) \tag{29}$$

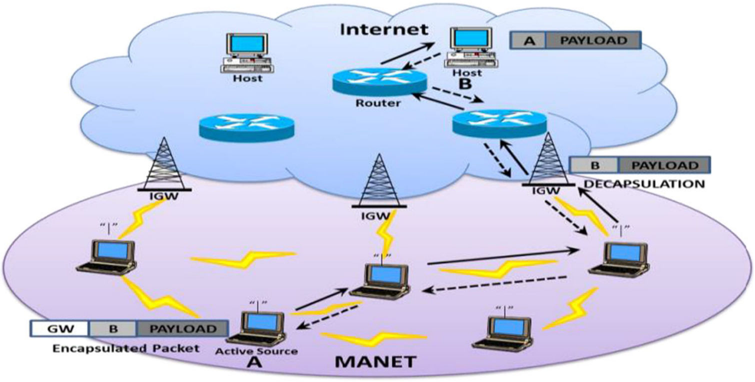
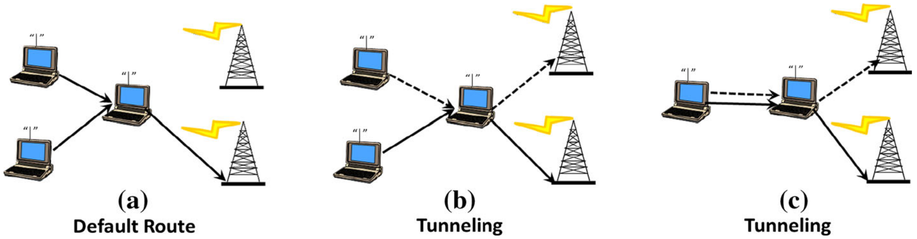


Fig. 5 Half-tunneling forwarding strategy


(a) Default Route

(b) Tunneling

(c) Tunneling

Fig. 6 **a** A default route points to only one IGW and aggregates all traffic. **b** and **c** With tunneling two MNs can share an intermediate node while still maintaining tunnels to different IGWs and forward traffic for one destination over several IGWs at once (multi-homing)

Table 2 Notations used in the analytical model

Notation	Meaning
N	Total number of nodes covering a certain area
N_G	Number of IGWs
N_{adhoc}	Number of ad hoc nodes $(N - N_G)$
$\lambda_{GA} \times t$	Transmission rate of the GWADVs during a time interval t
$\lambda_S \times t$	Mean number of IGW discoveries that need to be done during a time interval t
N_S	Number of AS nodes
S_{RQP}	Sum of IGW route requests and replies
N_{ARQ}	Number of accepted requests by each IGW
M_{PL}	Mean path length
N_{TTL}	Number of MNs in the TTL range for each IGW
N_{HY}	Number of ASs out the TTL range
N_{PRO}	Number of ASs within the TTL range
N_{GN}	Number of MNs, one hop away from the IGW
N_{SN}	Number of MNs, one hop away from the AS
D_N	The mean network diameter

4.2 Reactive IGW Discovery Overhead

In the reactive IGW discovery approach, whenever an MN wants to access the Internet, it broadcasts a GWSOL message. In response, each IGW sends a special reply message back to the AS node as much as the accepted requests, so that the AS can set up a multi-path route to each IGW. The overhead Θ_R of the reactive protocol can be computed as follows:

$$\Theta_R = S_{RQP} \times \sum_{a=1}^{N_S} \lambda_S \times t \tag{30}$$

$$S_{RQP} = N_{adhoc} + \sum_{b=1}^{N_G} N_{ARQ} \times M_{PL} \tag{31}$$

4.3 Hybrid IGW Discovery Overhead

In the hybrid approach, ASs in the TTL range behaves proactively, and those beyond that range search for the IGWs reactively. Thus, the hybrid approach has an overhead which is a combination of the reactive and proactive protocols. Thus, the overhead Θ_H is computed as follows:

$$\Theta_H = N_{TTL} \times N_G \times (\lambda_{GA} \times t) + S_{RQP} \times \sum_{c=1}^{N_{HY}} \lambda_S \cdot t \tag{32}$$

$$N_{HY} = N_S \times \left(1 - \frac{N_{PRO}}{N_S}\right) \tag{33}$$

where $\boldsymbol{N_{PRO}/N_S}$ is the probability that the ASs resides in the proactive area and $\boldsymbol{N_{TTL}}$ for this approach is the number of MNs in the scope of TTL hops.

4.4 The Proposed ADMIGW Discovery Overhead

In the proposed ADMIGW approach, the TTL range is not constant and determined according to the current network traffic conditions. The overhead of the proposed ADMIGW approach is computed as follows:

$$\Theta_A = (\lambda_{GA} \times t) \times \sum_{b=1}^{N_G} N_{TTL} + S_{RQP} \times \sum_{c=1}^{N_{HY}} \lambda_S \times t \tag{34}$$

$$N_{TTL} \geq N_{GN} \times D_N \tag{35}$$

$$S_{RQP} \leq N_{SN} \times M_{PL} \times \sum_{b=1}^{N_G} N_{ARQ} \cdot M_{PL}. \tag{36}$$

The analytical comparison of the four approaches is seen in Fig. 7. Assuming that there are 60 N_{adhoc} nodes in a square lattice covering a certain area and, the TTL value of the hybrid protocol is set to 2, λ_{GA} is set to 1/5 and the time t is set to 300 s. In Fig. 7a, the N_G was set to 4 nodes positioned in the corners of the lattice and in Fig. 7b, the N_S was set to 15 source nodes.

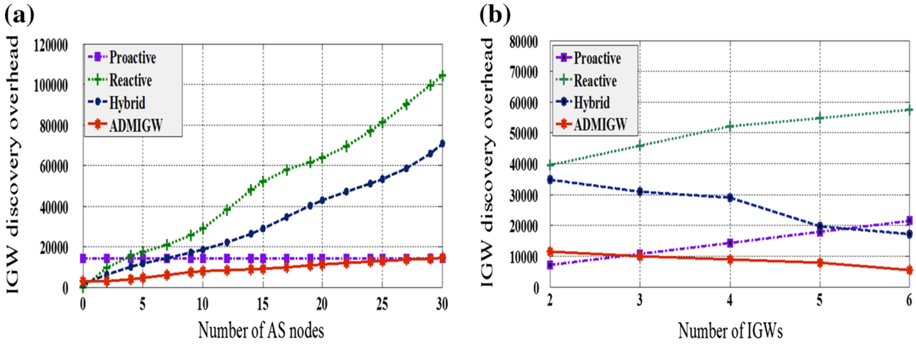


Fig. 7 IGW discovery overhead for the analytical model: **a** under various numbers of ASs and **b** under various numbers of IGWs

As shown, in the proposed ADMIGW, the proposed IGWs advertising procedure enables almost all ASs to be covered with the GWADVs without generating an excessive overhead compared to the proactive approach. On the other hand, the ASs that cannot receive the GWADV messages can solicit the IGW information using the proposed GWSOL algorithm which is also consumes a significantly smaller overhead than the reactive approach. Thus, the proposed ADMIGW incurs a good scalability both with respect to the number of ASs and IGWs. Although the proposed ADMIGW provokes a higher overhead than the proactive approach when there are a few IGWs, it gets the best values for the remaining cases.

5 Implementation and Performance Evaluation

In this section, a brief description of the simulation scenario, mobility and communication model, performance evaluation metrics, and finally the simulation results are presented.

5.1 Simulation Environment

5.1.1 Simulation Scenarios

The performance of the proposed framework is evaluated through simulations using NS-2 [45]. A network of 60 MNs randomly distributed in a square area of 2000 × 2000 m^2 is considered with four IGWs and radio transmission range of 250 m. In the simulated scenarios, the MNs are randomly portioned into set of ASs; the number of AS nodes differs to model different network traffic load. In the hybrid approach, all IGWs use a TTL value of two for the GWADVs messages.

5.1.2 Mobility and Communication Model

The used mobility model is based on, the random waypoint mobility model [46]. Each MN stays in one location for a certain pause time period (set to 10 s), once this time expires the MN chooses a random destination in the simulation area and a random speed, between a minimum speed of 0 m/s and a maximum speed of 30 m/s, and travels toward the newly chosen destination with the selected speed. Upon arrival, the MNs pause for a certain time period before starting the process again. This process repeated throughout the simulation, causing continuous changes in the topology of the underlying network. Simulations are run for 300 s. The simulation model parameters employed in the study are summarized in Table 3.

5.1.3 Performance Evaluation Metrics

The performance of the proposed framework is analyzed under various traffic loads and maximum speed with respect to the following performance metrics:

- *Average end-to-end delay* Is the average time taken for a packet to be transmitted from the AS to the IGW nodes. It is calculated as the difference between the time the packet is received by the IGW and the time the packet is generated by the AS.
- *Packet delivery ratio* Is defined as the ratio of successfully received data packets to the total number of the generated data packets.

Table 3 The simulation parameters

Parameter	Value
Simulation area	2000×2000 m^2
Transmission range	250 m
Bandwidth	2 Mb/s
Simulation time	300 s
Number of mobile nodes	60 nodes
Number of source nodes	5–30 AS nodes
Number of IGWs	4 IGWs
Traffic type	CBR (Constant bit rate)
Packet size	512 bytes
Packet sending rate	5 packets/s
Arrival rate	150 kbps
A_g	5 agents
t_{NOM}	1 s
Pause time	10 s
Mobility model	Random waypoint mobility
Node speed	0–30 m/s
Periodic GWADVs interval (T)	5 s
$a_1 = a_2 = a_3 = a_4$	0.25
I_{Min} and I_{Max} (*from simulation*)	3 and 9 s; respectively
Th_{Min} and Th_{Max} (from simulation)	−0.6 and 0.7; respectively
TTL value of the hybrid approach	2 hops
$\alpha = \beta = \gamma$	1
w_1, w_2, w_3	$w_1 = 0.5$, $w_2 = 0.3$, $w_3 = 0.2$

- *Normalized routing overhead* Is defined as the total number of control packets divided by the number of successfully delivered data packets during the simulation time.
- *Load-balancing* Is the effectiveness of load-balancing among IGWs. It is measured by the ratio of the heaviest traffic load to the lowest traffic load among all IGWs.

5.2 Simulation Results

5.2.1 Demonstrate the Effectiveness of Increasing the Number of MNs on the Delay

In order to prove the efficiency of the proposed framework and also to determine the suitable number of IGWs that can cope with the number of MNs, the average end to end delay and the packet delivery ratio are examined while gradually varying the number of MNs and IGWs from 50 to 300 and from 2 to 12; respectively, under various traffic loads and maximum speed of 30 m/s.

As shown in Fig. 8, the average end to end delay and the packet delivery ratio are significantly degraded with the increase of the number of MNs while keeping the number of IGWs low. This is because, lower number of IGWs mean high time for both reaching the IGWs, there will be a high traffic load around the area of each IGW and almost all traffic goes through specific directions, and processing time at each IGW. Also, IGWs cannot serve all incoming traffic, which leads to degrading the packet delivery ratio.

On the other hand, increasing the number of IGWs may help the network to cope with all traffic in terms of delay and delivery ratio. However, in each case of the number of MNs, when the number of IGWs increased from a specific value, the performance is degraded. This is because many IGWs mean many GWADVs overhead messages which in turn may cause many collisions and congestions.

As shown in Fig. 8, four IGWs can cope with a network consists of 50–100 MNs in terms of delay and packet delivery ratio. Also, as shown, the proposed framework efficiently can cope with the delay and delivery ratio even with the increasing number of MNs, when a suitable number of IGWs are used for each case.

5.2.2 Demonstrate the Effectiveness of the GWADV Interval

The effectiveness of the GWADV interval measurement is demonstrated by examining the packet delivery ratio, average delay and normalized routing overhead while gradually varying the IGW time interval from 1 to 11. First, to adjust the I_{Max} and I_{Min} values for the proposed feedback control system. Also, so as to verify a fair successive comparison between proactive, hybrid and ADMIGW protocols, the appropriate GWADV interval that achieve acceptable average delay and normalized overhead while not sacrificing connectivity for the proactive and hybrid protocols is determined based on the proposed simulation environment with varying speed and numbers of AS nodes.

Figure 9 illustrates the impact of varying the periodical GWADVs interval with different offered load. As shown in the proactive approach, the normalized overhead increases dramatically as the GWADVs interval decreases. However, the increasement of the GWADVs interval causes a sharp increase in the average delay and a great packet drop off specifically when the GWADV interval is greater than 7 s. On the other hand, the hybrid approach achieves the lowest overhead; however, it has the highest delay and lowest delivery ratio compared to others.

As shown, the proposed ADMIGW achieves the lowest delay and the highest delivery ratio, while achieving a moderate routing control overhead. From these simulation results, the I_{Min} and I_{Max} are set to 3 and 9 s; respectively. Also, a GWADVs interval of 5 s shows an acceptable normalized overhead, average delay, and connectivity for the proactive and hybrid protocols, thus in the subsequent simulations, the periodical GWADVs interval will set to 5 s.

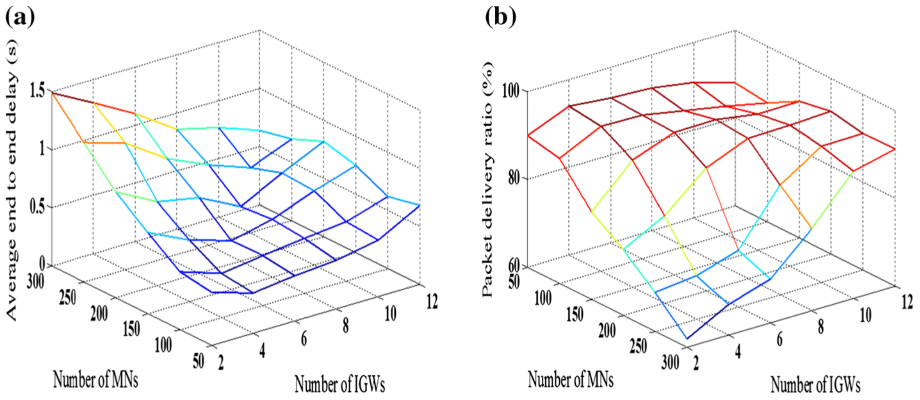


Fig. 8 Effects of varying number of MNs and IGWs in different traffic load: **a** average end-to-end delay and **b** packet delivery ratio

5.2.3 Determine the Optimum Maximum and Minimum Threshold Values

The maximum and minimum threshold values (Th_{Max} and Th_{Min}), used in the proposed feedback control system, should be selected very carefully in order to not scarcely and/or gratuitously cover the network with the GWADV messages. The Th_{Max} and Th_{Min} are heuristically determined according to an exhaustive set of simulations. The traffic load and maximum speed are varied to model different amount of network status changes (ΔI). The three dimensional graphs shown in Fig. 10 depict the obtained average end to end delay, packet delivery ratio and normalized routing overhead as a function of the Th_{Max} and Th_{Min} values under various traffic loads and maximum speed, which leads to changes for ΔI, every point in these graphs represents the mean value of 10^2 simulations.

As shown, setting the Th_{Max} to a low value leads to a sharp decrease in the network performance in terms of delay and delivery ratio, while achieving a low overhead, this is because of the high probability of positive ΔI to be greater than Th_{Max} which in turn leads to setting the next interval to the maximum interval, although the MNs is actually needed the IGWs information. In the same way, setting the Th_{Min} to a high value leads to improving the network performance in terms of delay and delivery ratio, however, it consumes a large overhead, this is because, there will be a high probability of negative ΔI to be smaller than the Th_{Min} which in turn leads to gratuitously dispatch the GWADVs after the minimum interval.

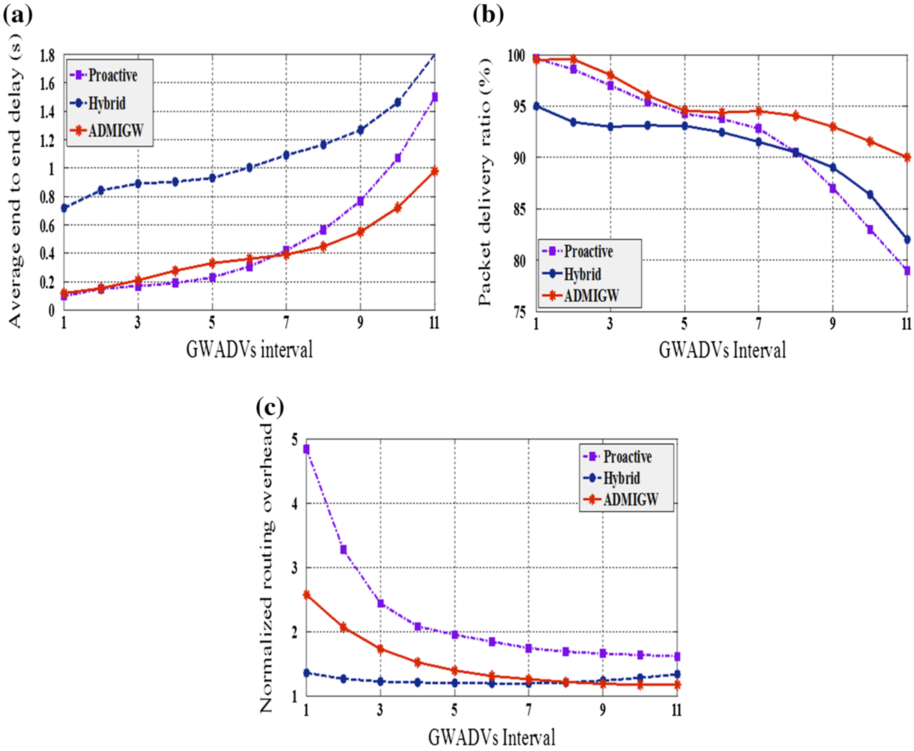


Fig. 9 Effects of varying IGWs interval in different traffic load: **a** average end-to-end delay, **b** packet delivery ratio, and **c** normalized routing overhead

As shown in Fig. 10, the simulations show that, when the positive amount of network status change ΔI is greater than or equal to 70 %, the network is highly stable and thus the interval can be set to maximum in order to preserve the overhead. On the other hand, when the negative amount of ΔI is lower than 60 %, the network is in critical need for the IGWs information, thus the interval must be set to the minimum. Thus, in this paper the Th_{Max} is set to 0.7 and Th_{Min} is set to -0.6.

5.2.4 Adjust Values of the Weighting Factors Used in the IGW Selection Mechanism

As discussed earlier in Sect. 3.3, three different weighting factors w_1, w_2, w_3 are assigned for the TT, GL, and HC metrics; respectively. In determining the suitable value of each weighting factor, a comparison of the performance of each metric is evaluated separately in terms of load-balancing, the average end-to-end delay, and packet delivery ratio under various traffic load. Afterward the weighting factor of each metric is set according to the importance of this metric which is derived from the results; finally the optimum performance result can eventually be obtained. In Fig. 11, four different scenarios are analyzed, as depicted in Table 4.

When an IGW is handling a heavy traffic, AS node should start searching for an alternate IGW to achieve a better load-balancing among all IGWs. As shown in Fig. 11a, the IGW traffic load based mechanism effectively distributes the traffic so as to avoid the situation where an IGW becomes a traffic bottleneck. On the other hand, the load-balancing is degraded in the two other mechanisms, especially with the minimal hops based mechanism.

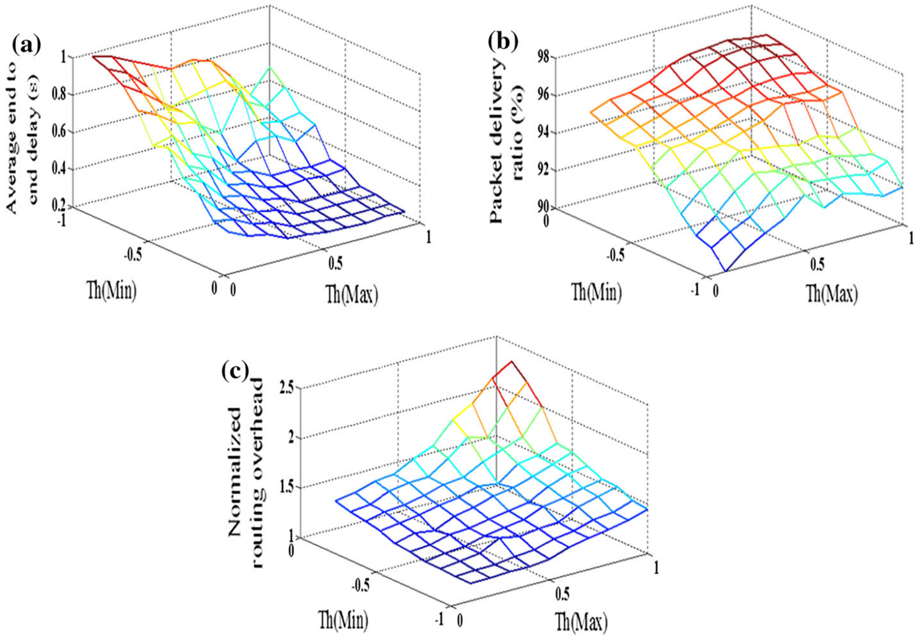


Fig. 10 Effects of varying maximum and minimum threshold values on: **a** average end to end delay, **b** packet delivery ratio and **c** normalized routing overhead

Figures 11b, c show that, among the three mechanisms, the path quality based mechanism has the lowest average delay and the highest packet delivery ratio under the most cases of different network traffic load. That is because it selects the IGW with the lowest TT metric, i.e. a good quality path, which also in some cases reflects the low traffic load of the related IGW. The minimal hops based mechanism gets the poorest end to end delay and packet delivery ratio, since it simply selects an IGW with the shortest path, not considering their related path quality and/or their density. Although the GL metric performs better than the minimal path selection, it will not end up using a good IGW in all cases, since the metric only considers offering load to select an IGW. Shortly, an IGW selection using a single metric does not perform well. So, considering multiple QoS metrics for IGW selection can achieve a better performance.

From the results, it is clear that the TT metric is the most important metric, subsequently, the GL metric, and finally the HC metric. Thus, for the proposed IGW selection mechanism, the three metrics are combined with the following weighting factors: $w_1 = 0.5$, $w_2 = 0.3$ and $w_3 = 0.2$ for TT, GL, and HC metrics; respectively. As shown, the superiority of the proposed IGW selection mechanism becomes apparent, especially under heavy traffic. It achieves a good load-balancing because it considers the GL metric and packet distribution. Also, it achieves the best average end-to-end delay and packet delivery ratio because it combines the three QoS metrics, so it selects a lightly loaded IGW which also have a good quality and short path from the AS node.

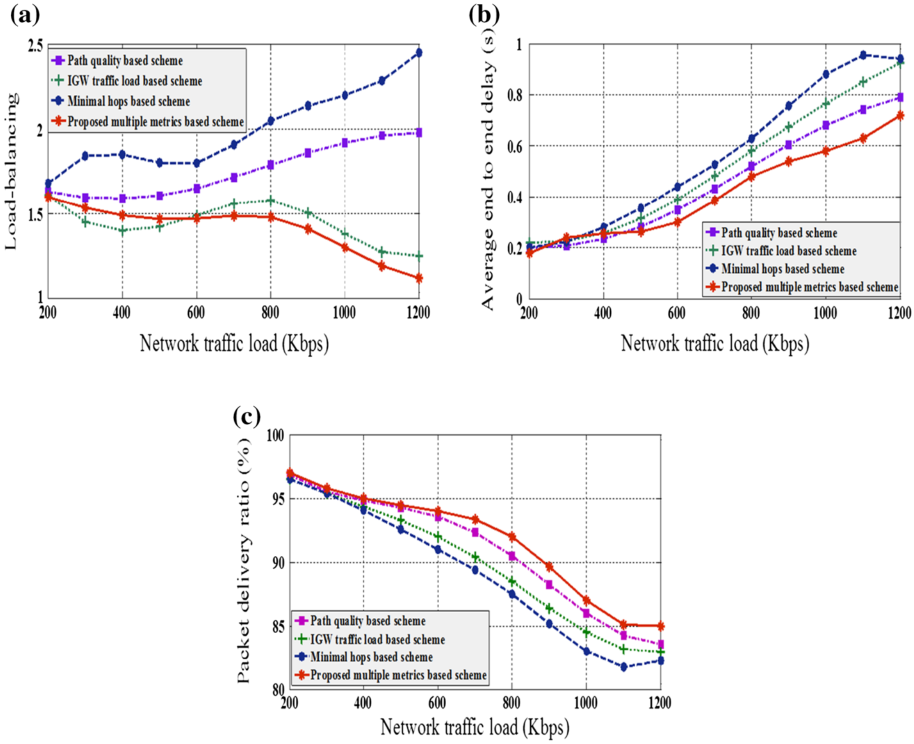


Fig. 11 Performance evaluation with different traffic load for access IGW selection: **a** load-balancing, **b** average end-to-end delay, and **c** packet delivery ratio

5.2.5 Performance Evaluation

In this section, the performance evaluation of the proposed ADMIGW is first compared with the basic proactive, reactive and hybrid IGWs discovery approaches. Afterward, it compared with other adaptive hybrid approaches.

5.2.5.1 Performance Evaluation of ADMIGW and Other IGW Discovery Approaches Figures 12 and 13 show the simulation results for the proactive, reactive, hybrid and ADMIGW approaches with varying number of AS nodes and maximum speed; respectively.

Figures 12 and 13 conclude that the proactive IGW discovery approach can achieve good connectivity, but also heavily increment the overhead. Such overhead is unnecessary in the situation when there are few ASs desired for Internet access. On the other hand, the reactive connectivity ties the overhead of maintaining connectivity to external traffic patterns, so MANET's scarce resources are not burdened with unnecessary overhead when the external traffic is small, but it scales poorly regarding the number of ASs willing to access the Internet. The hybrid approach combines the advantages of both proactive and reactive discovery approaches to balance the delay and control overhead. The proposed ADMIGW achieves almost the lowest packet delay and the highest delivery ratio compared to the other approaches, while consuming a much lower overhead than the proactive approach. Thus, the designed ADMIGW discovery protocol scales well with network size and traffic load.

5.2.5.2 Performance Comparison of ADMIGW and Other Adaptive IGW Discovery Approaches To verify the goodness of the proposed framework, it will be compared with three other adaptive IGW discovery approaches:

- *Yuste et al.* [21, 22] Proposed a type-2 fuzzy logic system protocol, which installed in the MNs to estimate the stability of the routes in a distributed manner without requiring any new control message. The main idea here is that the GWADVs messages should only be transmitted through routes with a high probability of enduring at least until the next GWADV is generated. The MN that receives the GWADV decides to rebroadcast it or not according to its own fuzzy output and its relaying status, i.e. if it is relaying data of an AS towards the IGW, its fuzzy output is compared with a minimum threshold value set to 0.3, otherwise, the output is compared with a high threshold value set to 0.8. In both cases, if the output exceeds the threshold value the MN forward the GWADV message, otherwise, it discards it.

Table 4 IGW selection scenarios

Scenario	w_1	w_2	w_3	Note
Path quality based IGW selection	1	0	0	Based solely upon the TT metric
IGW traffic load based IGW selection	0	1	0	Totally depends on the GL metric
Minimal hops based IGW selection	0	0	1	Totally depends on the HC metric
Proposed multiple metrics based IGW selection	0.5	0.3	0.2	Combines the three metrics based on their importance according to the obtained results

- *Javaid et al.* [24] Proposed an adaptive distributed IGW discovery approach. The main idea here is that the GWADVs are targeted only to those nodes looking for the IGWs and other nodes should not be hampered with the periodic GWADVs. Initially the IGWs sends the GWADV message with TTL = 1. Then, on the reception of the GWADV message, only the relaying MNs forward it further, which in turn result into the formation of an active region comprises of all the MNs between the IGWs and the ASs.
- *Lin et al.* [23] Proposed an adaptive IGW discovery approach where unidirectional links are removed from route computations. The main idea here is the dynamic adjusting of the broadcast range and the sending interval of the GWADVs in terms of the network conditions. Each IGW keeps an ASs list while setting a lifetime for each entry to remove ASs departs from the IGW or never needs the Internet service. The TTL range is computed dynamically by each IGW according to the close or far of the ASs for obtaining a mean TTL value.

The three protocols considered AODV [16] as the support routing algorithm between the MANET and the Internet. Figures 14 and 15 show the performance comparison of the proposed ADMIGW protocol and other adaptive IGW discovery protocols discussed above in terms of the packet delivery ratio, the average end-to-end delay and normalized routing overhead with varying number of AS nodes and maximum speed; respectively.

In Yuste et al.'s protocol, the transmission of the GWADVs only to the stable routes makes it an efficient protocol, however, MNs require high processing power consumption

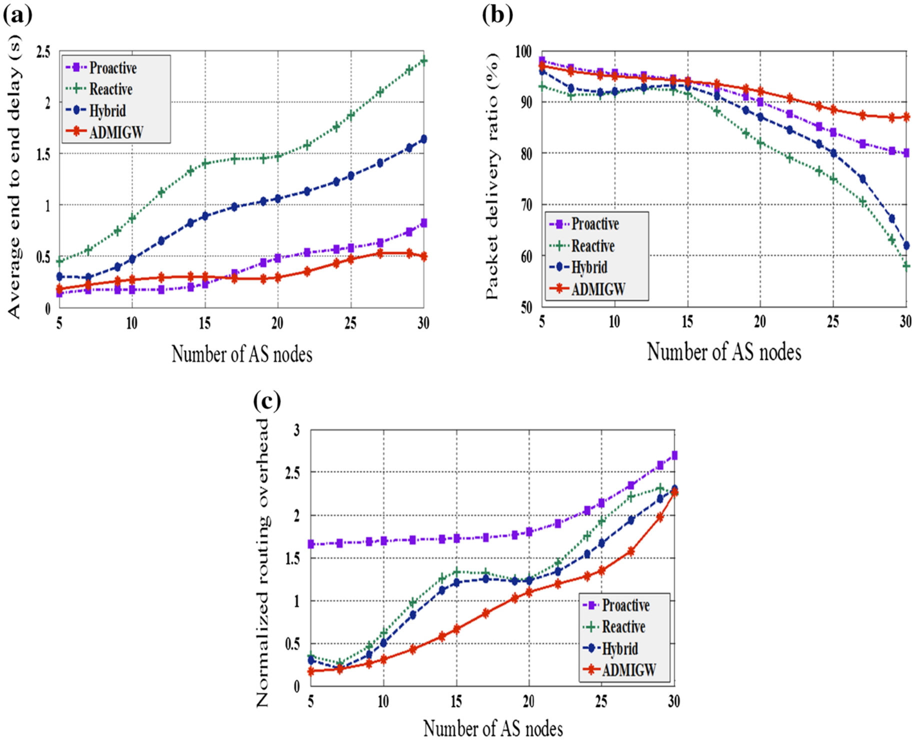


Fig. 12 Performance evaluation with varied number of ASs: **a** average end-to-end delay, **b** packet delivery ratio, and **c** normalized routing overhead

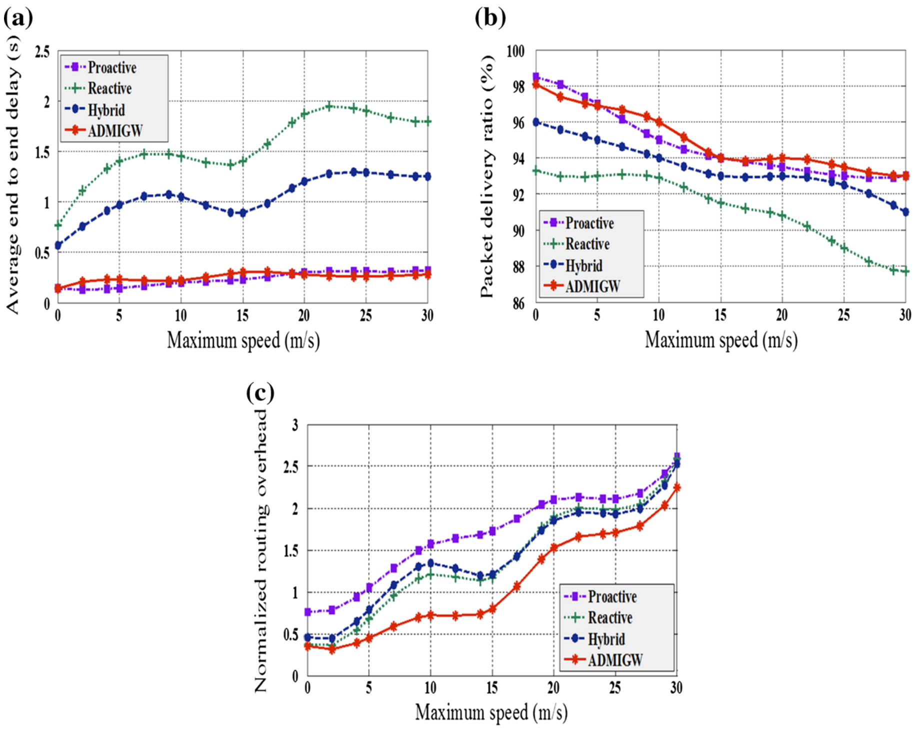


Fig. 13 Performance evaluation with varied maximum speed: **a** average end-to-end delay, **b** packet delivery ratio, and **c** normalized routing overhead

for the availability of its forwarding status, i.e. either to forward the GWADVs or not, all the times. Also, in the fuzzy system, the determination of specified threshold values degrades its performance, as shown in Figs. 14 and 15a, b the average delay and delivery ratio are degraded with the increasing of the number and speed of the ASs due to the sharp rules used in the fuzzy system which prevent forwarding the GWADVs messages and make the protocol reacts nearly as the reactive protocol. Also, when the network is somewhat stable, the normalized routing overhead becomes nearly similar as the proactive protocol because almost all MNs can forward the GWADVs as shown in Figs. 14 and 15c.

In Javaid et al.'s protocol, the formation of the active regions may help the ASs to quickly adapt to the continuously changing network topology in case of high mobility. However, the active regions formation consumes an excessive overhead when the network is somewhat stable, especially with the increase of the number of ASs as shown in Figs. 14 and 15c. Also, the TTL value is estimated only after the arrival of the last data packet for each connection, thus the decision to make is very dependent on a single event which in turn degrades the performance, as shown in Figs. 14 and 15a, b.

In Lin et al.'s protocol, the dynamic change of the TTL value according to the network conditions seems to be a good idea, however, when the TTL value reaches a certain level, substantial congestion and packet collisions may arise which in turn degrades the performance. Also, unlike other adaptive approaches, IGWs periodically search the AS list to obtain the distributing of ASs around it, i.e. the adoption is performed centrally at the

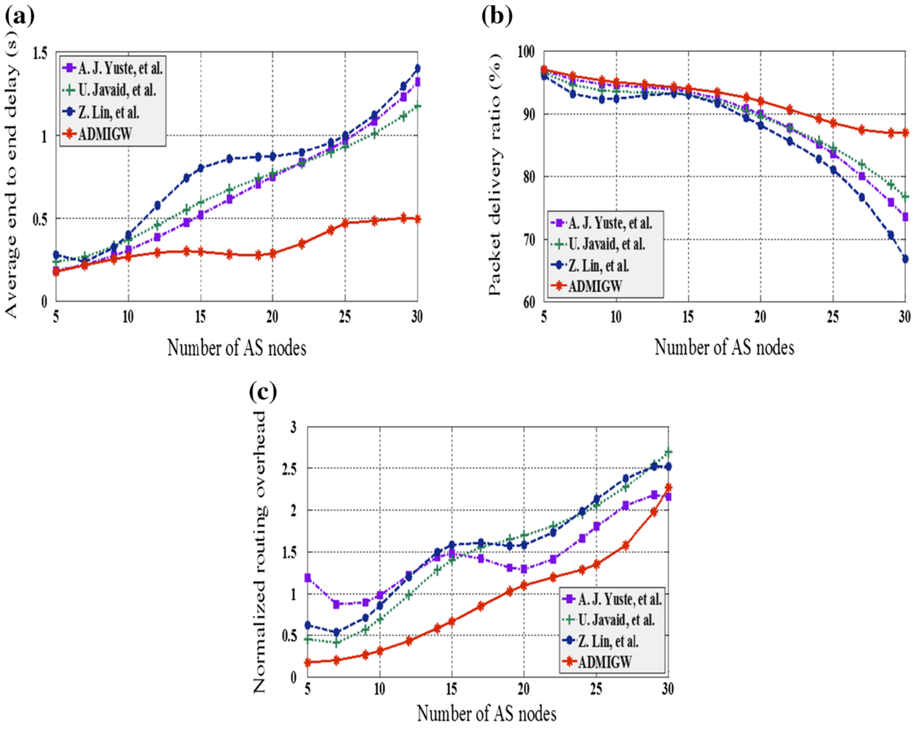


Fig. 14 Performance comparison with varied number of ASs: **a** average end-to-end delay, **b** packet delivery ratio, and **c** normalized routing overhead

IGWs and this of course will affect the efficiency of them. Thus, this protocol brings the lowest performance, in most cases, among all other protocols as shown in Figs. 14 and 15.

In Figs. 14 and 15, the advantages of the proposed ADMIGW approach become apparent, especially under heavy traffic and high mobility cases. It outperforms other adaptive approaches in terms of delay by about 10–30 % and delivery ratio by about 9–18 %. The hybrid routing algorithm used for the IGWs solicitation procedure outperforms the traditional AODV algorithm used in the above adaptive protocols because of the cooperation of the RFAs and PFAs which continuously establish, maintain and improve multiple and good quality paths between the AS nodes and the IGWs. Additionally, the proposed dynamic and totally distributed IGWs advertising protocol enables not only almost all NMNs but also other, non-NMNs to be covered adaptively, with the help of the optimized feedback control system, with the GWADVs messages which in turn, helps the MNs to quickly learn the route towards the IGWs without generating an excessive overhead, i.e. keeping overhead costs low.

6 Conclusion

In this paper, an efficient and complete framework that comprises the main phases of the solution for MANET–Internet integration is proposed. The proposed framework comprises a new, efficient and adaptive multipath gateway discovery protocol for MANETs–Internet

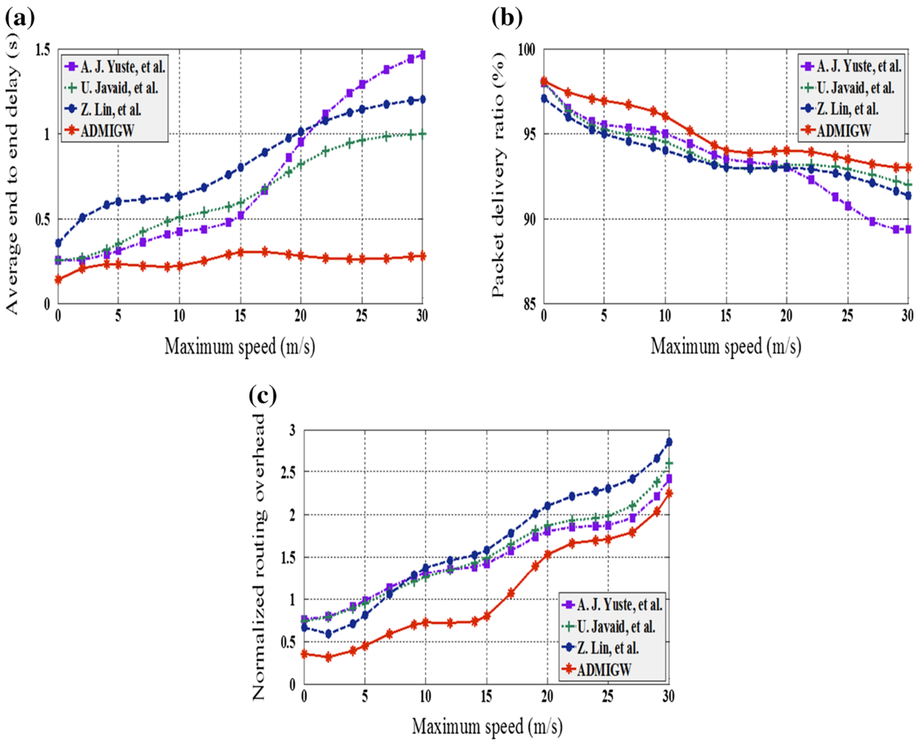


Fig. 15 Performance comparison with varied maximum speed: **a** average end-to-end delay, **b** packet delivery ratio, and **c** normalized routing overhead

integration inspired by the ant colony optimization (ACO) algorithms. In addition, a new proactive area adjustment protocol that adapts its behavior, based on the present network conditions is proposed to provide efficient and faster discovery of the gateways. The gateway advertisement time interval is adjusted according to a new optimized feedback control system. Also, an improved QoS-based gateway selection mechanism which reduces the data drop rate and provides load-balancing across a set of the discovered access gateways is proposed. The simulation results proved the effectiveness of the proposed framework compared to the already existing approaches. The proposed framework has a great impact on improving the delay and the delivery ratio, while attaining tolerable overhead and fair load distribution among all gateways.

References

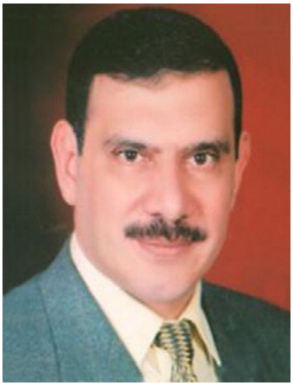
1. Conti, M., & Giordano, S. (2014). Mobile ad hoc networking: milestones, challenges, and new research directions. *IEEE Communications Magazine, 52*(1), 85–96.
2. Boukerche, A., Turgut, B., Aydin, N., Ahmad, M. Z., Boloni, L., & Turgut, D. (2011). Routing protocols in ad hoc networks: A survey. *ScienceDirect, Computer Networks, 55*, 3032–3080.
3. Kostin, A., Oz, G., & Haci, H. (2014). Performance study of a wireless mobile ad hoc network with orientation-dependent internode communication scheme. *International Journal of Communication Systems, 27*, 322–340.

4. Chen, J., Yuan, Z., & Wang, L. (2014). An apropos signal report and adaptive period (ASAP) scheme for fast handover in the fourth-generation wireless networks. *ScienceDirect, Journal of Network and Computer Applications, 45*, 15–26.
5. Boudriga, N., Obaidat, M. S., & Zarai, F. (2008). Intelligent network functionalities in wireless 4G networks: Integration scheme and simulation analysis. *ScienceDirect, Computer Communications, 31*, 3752–3759.
6. Attia, R., Rizk, R., & Ali, H. A. (2015). Internet connectivity for mobile ad hoc network: A survey based study. *Wireless Networks*,. doi:10.1007/s11276-015-0922-3.
7. Fekri, M., & Shrikant, K. (2007). A survey of integrating IP mobility protocols and mobile ad hoc networks. *IEEE Communications Surveys & Tutorials, 9*(1), 14.
8. Ruiz, P. M., Ros, F. J., & Gomez-Skarmeta, A. (2005). Internet connectivity for mobile ad hoc networks: Solutions and challenges. *IEEE Communications Magazine, 43*, 118–125.
9. Nordstrom, E., Gunningberg, P., & Tschudin, C. (2011). Robust and flexible Internet connectivity for mobile ad hoc networks. *ScienceDirect, Ad Hoc Networks, 9*, 1–15.
10. Wang, Y.-L., Song, M., Wei, Y.-F., Wang, Y.-H., & Wang, X.-J. (2014). Improved ant colony-based multi-constrained QoS energy-saving routing and throughput optimization in wireless Ad hoc networks. *ScienceDirect, The Journal of China Universities of Posts and Telecommunications, 21*, 43–53.
11. Caria, D. C., & Godbole, V. V. (2013). New approach for routing in mobile ad-hoc networks based on ant colony optimisation with global positioning system. *IET Networks, 2*, 171–180.
12. Attia, R., Rizk, R., & Mariee, M. (2010). An ant inspired QoS routing algorithm for MANETs. *The International Journal of Ad Hoc & Sensor Wireless Networks, 10*(2–3), 111–134.
13. Sandoval, E. I., Galvan, C. E., & Galvan-Tejada, J. I. (2012). Multicast routing and interoperability between wired and wireless ad hoc network. *ScienceDirect, Procedia Engineering, 35*, 109–117.
14. Al-Surmi, I., Othman, M., & Ali, B. M. (2012). Mobility management for IP-based next generation mobile networks: Review, challenge and perspective. *ScienceDirect, Journal of Network and Computer Applications, 35*, 295–315.
15. Domingo, M. C., & Remondo, D. (2008). QoS support between ad hoc networks and fixed IP networks. *ScienceDirect, Computer Communications, 31*, 2646–2655.
16. Hamidian, A., Korner, U., & Nilsson, A. (2005). Performance of internet access solutions in mobile ad hoc networks. In *Springer's lecture notes in computer science* (*LNCS*), pp. 189–201.
17. Majumder, K., Ray, S., & Sarkar, S. K. (2011). Design and analysis of the gateway discovery approaches in MANET. *High Performance Architecture and Grid Computing Communications in Computer and Information Science, 169*, 397–405.
18. Shimizu, M., & Takami, K. (2014). Improving communication quality by considering route stability for inter-gateway mobile ad-hoc networks. In *International conference on information and communication technology convergence (ICTC)*, pp. 142–147.
19. Majumder, S., & Asaduzzaman. (2014). A hybrid gateway discovery method for mobile ad hoc networks. In *3rd international conference on informatics, electronics & vision (ICIEV)*, pp. 1–6.
20. Palani, K., & Ramamoorthy, P. (2014) Performance evaluation of QoS based DSDV protocol using an integration approach for hybrid networks. In *International conference on green computing communication and electrical engineering (ICGCCEE)*, pp. 1–6.
21. Yuste, A. J., Trivino, A., & Casilari, E. (2013). Type-2 fuzzy decision support system to optimise MANET integration into infrastructure-based wireless systems. *ScienceDirect, Expert Systems with Applications, 40*, 2552–2567.
22. Yuste, A. J., Trivino, A., Casilari, E., & Trujillo, F. D. (2011). Adaptive gateway discovery for mobile ad hoc networks based on the characterization of the link lifetime. *IET Communications, 5*, 2241–2249.
23. Lin, Z., Yuan-an, L., Kai-ming, L., Lin-bo, Z., & Ming, Y. (2010). An adaptive algorithm for connecting mobile ad hoc network to Internet with unidirectional links supported. *ScienceDirect, The Journal of China Universities of Posts and Telecommunications, 17*, 44–49.
24. Javaid, U., Rasheed, T., Meddour, D.-E., & Ahmed, T. (2008). Adaptive distributed gateway discovery in hybrid wireless networks. In *Proceedings of the IEEE wireless communications and networking conference (WCNC)*, pp. 2735–2740.
25. Domingo, M. C., & Prior, R. (2007). An adaptive gateway discovery algorithm to support QoS when providing Internet access to mobile ad hoc networks. *Journal of Networks, 2*(2), 33–44.
26. Park, B.-N., Lee, W., & Lee, C. (2007). QoS-aware internet access schemes for wireless mobile ad hoc networks. *ScienceDirect, Computer Communications, 30*, 369–384.
27. Zhong, S., & Zhang, Y. (2013). How to select optimal gateway in multi-domain wireless networks: alternative solutions without learning. *IEEE Transactions on Wireless Communications, 12*, 5620–5630.

28. Bouk, S. H., Sasase, I., Ahmed, S. H., & Javaid, N. (2012). Gateway discovery algorithm based on multiple QoS path parameters between mobile node and gateway node. *Journal of Communications and Networks, 14*(4), 434.
29. Li, X., & Li, Z. (2010). A MANET accessing internet routing algorithm based on dynamic gateway adaptive selection. *Frontiers of Computer Science in China, 4*, 143–150.
30. Jaron, A., Pangalos, P., Mihailovic, A., & Aghvami, A. H. (2012). Proactive autonomic load uniformisation with mobility management for wireless internet protocol (IP) access networks. *IET Networks, 1*, 229–238.
31. Le-Trung, Q., Engelstad, P. E., Skeie, T., & Taherkordi, A. (2008). Load-balance of intra/inter-MANET traffic over multiple Internet gateways. In *Proceedings of the 6th international conference on advances in mobile computing and multimedia, MOMM'08*, pp. 50–57.
32. Wu, P., Cui, Y., Wu, J., Liu, J., & Metz, C. (2013). Transition from IPv4 to IPv6: A state-of-the-art survey. *IEEE Communications Surveys & Tutorials, 15*(3), 1407.
33. Wang, X., & Qian, H. (2015). Dynamic and hierarchical IPv6 address configuration for a mobile ad hoc network. *International Journal of Communication Systems, 28*, 127–146.
34. Grajzer, M., Zernicki, T., & Glabowski, M. (2014). ND ++—An extended IPv6 neighbor discovery protocol for enhanced stateless address autoconfiguration in MANETs. *International Journal of Communication Systems, 27*, 2269–2288.
35. Wang, X., & Qian, H. (2014). A tree-based address configuration for a MANET. *ScienceDirect, Pervasive and Mobile Computing, 12*, 122–137.
36. Yonghang, Y., Linlin, C., Chengping, T., & Hui, Z. (2012). A novel IP address auto-configuration scheme for MANET with multiple gateways. In *Proceedings of the 8th international conference on wireless communications, networking, and mobile computing (WiCOM)*, pp. 1–6.
37. Ancillotti, E., Bruno, R., Conti, M., & Pinizzotto, A. (2009). Dynamic address autoconfiguration in hybrid ad hoc networks. *ScienceDirect, Pervasive and Mobile Computing, 5*, 300–317.
38. Kim, D., Jeong, H.-J., Toh, C. K., & Oh, S. (2009). Passive duplicate address-detection schemes for on-demand routing protocols in mobile ad hoc networks. *IEEE Transactions on Vehicular Technology, 58*(7), 3558.
39. Cui, Y., Dong, J., Wu, P., Wu, J., Metz, C., Lee, Y. L., & Durand, A. (2013). Tunnel-based IPv6 transition. *IEEE Internet Computing, 17*, 62–68.
40. Xiaonan, W., & Shan, Z. (2014). Research on mobility handover for IPv6-based MANET. *Transactions on Emerging Telecommunications Technologies, 25*, 679–691.
41. Ding, S. (2009). Mobile IP handoffs among multiple Internet gateways in mobile ad hoc networks. *IET Communications, 3*, 752–763.
42. Tsumochit, J., Masaymatt, K., Ueharat, H., & Yokoymat, M. (2003). Impact of mobility metric on routing protocols for mobile ad hoc networks. In *Proceedings of IEEE pacific rim conference on communications, computers and signal processing (PACRIM)*, vol. 1, pp. 322–325.
43. Abramovici, A., & Chapsky, J. (2000). *Feedback control systems: A fast-track guide for scientists and engineers*. New York: Springer.
44. Venkata Rao, R. (2007). *Decision making in the manufacturing environment using graph theory and fuzzy multiple attribute decision making methods*. New York: Springer.
45. The network simulator—NS-2, http://www.isi.edu/nsnam/ns/.
46. Camp, T., Boleng, J., & Davies, V. (2002). A survey of mobility models for ad hoc network research. *Wireless Communications & Mobile Computing (WCMC): Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications, 2*(5), 483–502.

Dr. Radwa Attia is a staff member in the Electrical Engineering Department, Port Said University, Egypt. She received a B.Sc. and M.Sc. degrees in Computer and Control engineering, Suez Canal University, in 2004 and 2009; respectively. Her research interests are in the area of computer networking, including mobile networking, wireless networks, sensor networks, ad hoc networks, QoS, routing, traffic and congestion control, handoffs and cloud computing.

Dr. Rawya Rizk is an associate professor at the Electrical Engineering Department, Port Said University, Egypt. She received her B.Sc., M.Sc. and Ph.D. in Computer and Control Engineering from Suez Canal University in 1991, 1996 and 2001, respectively. Her research interests are in computer networking, including mobile networking, wireless networks, sensor networks, ad hoc networks, network security, QoS, routing, traffic and congestion control, handoffs and cloud computing. She is the Chief Information Officer (CIO), Port Said University.



Dr. Hesham Arafat Ali is a Professor in Computer Engineering and System and an associate Professor in Information System and Computer Engineering. He received a B.Sc. in Electronics Engineering, and M.Sc. and Ph.D. in Computer Engineering and Control from the Faculty of Engineering, Mansoura University, in 1986,1991 and 1997; respectively. He was an assistant professor at the University of Mansoura, Faculty of Computer Science in 1997–1999. From January 2000 up to September 2001, he joined as Visiting Professor to the Department of Computer Science, University of Connecticut. From 2002 to 2004 he was a vice dean for student affair the Faculty of Computer Science and Information University of Mansoura. He was awarded with the Highly Commended Award from Emerald Literati Club 2002 for his research on network security. He is a founder member of the IEEE SMC Society Technical Committee on Enterprise Information Systems (EIS). He has many book chapters published by international press and about 150 published papers in international (conf. and journal). He has served as a reviewer for many high quality journals, including Journal of Engineering Mansoura University. His interests are in the areas of network security, mobile agent, network management, search engine, pattern recognition, distributed databases, and performance analysis.