Elsevier Editorial System(tm) for Computer Communications
Manuscript Draft

Manuscript Number: COMCOM-D-13-00083

Title: CPCRA: A Contact-Predict Clustering-based Routing Algorithm in Large-scale Urban DTNs

Article Type: Special issue Opp-Net

Keywords: DTNs; routing; clustering; inter contact time; large-scale; simulation

Corresponding Author: Dr. Chunhe Xia, Doctor

Corresponding Author's Institution: Beihang University, Beijing 100191, P.R. China

First Author: Haiquan Wang, Doctor

Order of Authors: Haiquan Wang, Doctor; Wenjing Yang, Bachelor; Tao Zhu, Master; Ying Yang, Master; Chunhe Xia, Doctor

Abstract: In Delay Tolerant Networks (DTNs), with the increasing scale of network, single strategy routing to send more message copies at high overhead cost to improve the delivery rate and scalability. Achieving balance between the overall overhead and routing efficiency is difficult. Clustering has been proposed as an effective solution to improve the scalability of routing in traditional networks, e.g., MANET. Given the change of topology and intermittent connection in DTNs, how to adaptive cluster and limit the overall overhead including the clustering and routing overhead is a great challenge. Practically, distributed event-driven method can reduce clustering overhead and complexity by abandoning the global coordination of topology. In addition, contact is an important indicator of relationship between two nodes in DTNs, thus, it can be a good criterion when comparing similarities between two nodes. The distribution of inter-contact time (ICT) can infer contact probability in a period of time which ensures the stability of clustering in the future. In this paper, we propose a contact-predict clustering-based routing algorithm for large-scale urban DTNs. This algorithm computes the probability for each node pair and ensures that nodes in one cluster have higher contact strength. Depending on the required clustering method, intra-cluster and inter-cluster routing strategies will be chosen adaptively. Our simulations indicate that the proposed algorithm has better scalability for routing. The results demonstrate that clustering-based routing algorithm is rational and promising for large-scale urban DTNs.

Suggested Reviewers:

Cover Letter

Dear Editor,

I am submitting here a manuscript entitled “CPCRA: A Contact-Predict Clustering-based Routing Algorithm in Large-scale Urban DTNs”., which I would like to submit to Computer Communications. This paper is a substantial revision of the conference paper “Characterization and Modeling in Large-scale Urban DTNs”, which is published by the conference Local Computer Networks (LCN) 2012, Clearwater Beach,USA. The citation is

[1] Xia, Chunhe; Liang, Dong; Wang, Haiquan; Luo, Min; Lv, Weifeng; , "Characterization and modeling in large-scale urban DTNs," Local Computer Networks (LCN), 2012 IEEE 37th Conference on , vol., no., pp.352-359, 22-25 Oct. 2012.

The url is

http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6423647&isnumber=6423567.

“Characterization and Modeling in Large-scale Urban DTNs” focus more on the analysis and modeling on the urban city vehicles. We have made some improvements. Based on the clustering and contacting characters, we proposed a distributed clustering-based algorithm, and evaluate the algorithm of three aspects, that is, clustering, scalability and overhead.

Copy of the Abstract of “CPCRA: A Contact-Predict Clustering-based Routing Algorithm in Large-scale Urban DTNs” is shown as follows:

In Delay Tolerant Networks (DTNs), with the increasing scale of network, single strategy routing to send more message copies at high overhead cost to improve the delivery rate and scalability. Achieving balance between the overall overhead and routing efficiency is difficult. Clustering has been proposed as an effective solution to improve the scalability of routing in traditional networks, e.g., MANET. Given the change of topology and intermittent connection in DTNs, how to adaptive cluster and limit the overall overhead including the clustering and routing overhead is a great challenge. Practically, distributed event-driven method can reduce clustering overhead and complexity by abandoning the global coordination of topology. In addition, contact is an important indicator of relationship between two nodes in DTNs, thus, it can be a good criterion when comparing similarities between two nodes. The distribution of inter-contact time (ICT) can infer contact probability in a period of time which ensures the stability of clustering in the future. In this paper, we propose a contact-predict clustering-based routing algorithm for large-scale urban DTNs. This algorithm computes the probability for each node pair and ensures that nodes in one cluster have higher contact strength. Depending on the required clustering method, intra-cluster and inter-cluster routing strategies will be chosen adaptively. Our simulations indicate that the proposed algorithm has better scalability for routing. The results demonstrate that clustering-based routing algorithm is rational and promising for large-scale urban DTNs.

Corresponding Author:
Xia Chunhe,
State Key Laboratory of Virtual Reality Technology and System of
Beihang University, Beijing 100191, P.R. China
Telephone Number: 0086-13801312648

Email: xch@buaa.edu.cn

CPCRA: A Contact-Predict Clustering-based Routing Algorithm in Large-scale Urban DTNs

Wang Haiquan[234], Yang Wenjing[4], Zhu Tao[25], Yang Ying[25], Xia Chunhe[123*]

[1]State Key Laboratory of Virtual Reality Technology and System
[2]Beijing Key Laboratory of Network Technology
[3]School of Computer Science and Engineering
[4]School of Software
[5]School of Mathematics and system science
Beihang University, Beijing 100191, P.R. China
** The Corresponding Author*

Abstract: In Delay Tolerant Networks (DTNs), with the increasing scale of network, single strategy routing to send more message copies at high overhead cost to improve the delivery rate and scalability. Achieving balance between the overall overhead and routing efficiency is difficult. Clustering has been proposed as an effective solution to improve the scalability of routing in traditional networks, e.g., MANET. Given the change of topology and intermittent connection in DTNs, how to adaptive cluster and limit the overall overhead including the clustering and routing overhead is a great challenge. Practically, distributed event-driven method can reduce clustering overhead and complexity by abandoning the global coordination of topology. In addition, contact is an important indicator of relationship between two nodes in DTNs, thus, it can be a good criterion when comparing similarities between two nodes. The distribution of inter-contact time (ICT) can infer contact probability in a period of time which ensures the stability of clustering in the future. In this paper, we propose a contact-predict clustering-based routing algorithm for large-scale urban DTNs. This algorithm computes the probability for each node pair and ensures that nodes in one cluster have higher contact strength. Depending on the required clustering method, intra-cluster and inter-cluster routing strategies will be chosen adaptively. Our simulations indicate that the proposed algorithm has better scalability for routing. The results demonstrate that clustering-based routing algorithm is rational and promising for large-scale urban DTNs.

Key words: DTNs; routing; clustering; inter contact time; large-scale; simulation

1. INTRODUCTION

DTN is a branch of networks in where end-to-end linked paths intermittently connect. Nodes in this network move dynamically with strong randomness. The “storage-forward” mechanism of message transmission is no longer applicable. In many studies proposed “storage-carry-forward” mechanism[1-4] in protocols design of DTNs, e.g. Direct Delivery Routing[5], Max-probability Routing [6] and Epidemic Routing [7].Such protocols have their own special mechanism for message forwarding. However, they all execute only one type of strategy in the whole network. When the network scale increases, those protocols inevitably become unscalable and will require high overhead.

Clustering can efficiently improve the scalability of a routing protocol in traditional MANET applications [8-11]. In this method, nodes are clustered into different groups based on topology. In DTNs, move more frequently and randomly, which makes it a challenge to involve clustering in routing protocol design a challenge. Overhead may increase to maintain an overall clustering topology. Moreover, determining an effective criterion to describe similarity among nodes in the same cluster in such random networks is difficult. Some research focus to clustering in DTNs. In [12], a kind of DTN hierarchical routing algorithm was designed based on a mobility model where all nodes move according to strict and repetitive patterns. The algorithm is an efficient approach for a specific mobility model. However, it is limited by its difficulty to recognize the mobility character. In addition, applying the algorithm to a new network environment is difficult. The authors of [13]proposed a cluster-based forwarding algorithm based on graph theory in public transport networks. The overhead of overall information involved in creating the graph for a dynamic network in real time will increase greatly.

Contact [14-16]is an important feature in DTNs. Two nodes make contact when they appear in each other’s communication range, which indicates the opportunity to communicate. Inter-contact time (ICT) [17-19] between two nodes affects the transmission delay. To some degree, vehicles reveal their similarity through their contact pattern between them in large-scale urban DTNs, e.g. nodes that appear in the same area may contact more frequently, whereas nodes restricted to the same street may have contact for a longer period of time. If some inherent feature of contact in large-scale urban DTNs and the future contact probability between two mobile nodes can be predicted, contact can be a significant criterion for designing a clustering-based routing protocol in DTNs. In complicated network analysis and some DTNs research [18, 20] , historical contact frequency or duration is used as metric to describe contact probability, but these studies neglected the real mathematical relationship between the accurate contact probability and a priori knowledge or network characteristics. In[6, 21, 22], the authors proposed schemes for converting a priori knowledge into a form of value to predict contact probability. However, such schemes involve too many subjective parameter-setting procedures when formatting the expression of contact probability.

In this paper, large-scale urban DTNs are analyzed by gathering mobility traces from 12096 Beijing taxis for one week to reveal its inherent characteristics of clustering and contact. A clustering-based routing algorithm for such kind of DTNs is then designed. Our contributions are highlighted as follows:

- We analyze the distributions of active vehicles in Beijing. The results reveal that urban vehicles have aggregation character and generate some hotspots. Clustering is practical in such environment.
- Based on research, we believe the inter-contact time (ICT) follows an exponential distribution, which is applied to the taxi trace data in Beijing. Then, global fitting parameter settings are compared with individual parameter settings to set the exponential parameters. The result proves that the latter method, which involves setting different parameters for each pair of nodes based on their historical contacts, has certain accuracy. Therefore, the method is selected to forecast contact probability
- We propose a comprehensive clustering-based routing algorithm to improve scalability and suitability of routing in large-scale urban DTNs. Distributed event-driven strategy is involved in clustering algorithm design. Direct delivery and flooding strategy are adaptively executed in intra and inter cluster routing phases. Simulation results indicate that, compared with other classic routing protocols in DTNs, the proposed algorithm can guide the routing in a more scalable and adaptive way in large-scale urban DTNs.

The rest of this paper is organized as follows: Section 2 gives detailed analysis of clustering and contact property in large-scale urban DTNs. Section 3 proposes the clustering based routing algorithm in large-scale urban DTNs. Simulation and its results are presented in Section 4. Finally, we conclude our work and outline the future in Section 5.

2. CHARACTERISTICS OF LARGE-SCALE URBAN DTNS

In this section, we mainly focus on revealing the inherent characteristics of contact and clustering in large-scale urban DTNs, which we believe are essential properties that affect DTN protocol design and verification. The dataset we use to abstract network characteristics was collected from Beijing Urban Transportation Information System, which includes about 1.22×10^{8} records of 12096 taxis mobility traces in one week (2010.6.13 to 2010.6.19). Each vehicle is equipped with a unique GPS signal transmitter which reports the vehicle's status (location, velocity, direction) to data center every 15 seconds.

Based on abovementioned data, we analyzed the aggregation character and found that nodes tend to cluster. We therefore designed the algorithm to cluster based on contact probability. Two approaches are then compared to measure the contact strength. Finally, we choose that contact probability for each node pair is calculated which allow us to design a distributed algorithm to reduce the network overhead.

2.1 Clustering

We first quantitatively analyze the daily average vehicles density for four days including two workdays, shown in Figure 1(a) and Figure 1(b) and two weekends as shown in Figure 1(c) and Figure 1(d). We uniformly divide the whole network into 400 pieces of $1km^{2}$ each. The average vehicle density is computed in every separate grid. Beijing has a ring-like structure, which causes the cluster and uneven distribution of vehicles in central urban. In the central urban area, the average density of vehicles in one day is up to $197.4\,vehicles/km^{2}$ whereas that of remote areas reaches only $1\,vehicle/km^{2}$. A closer look at the distribution shows more fined-grain areas with high nodes aggregation, e.g. business centers, transportation hubs and airport. Areas with high nodes density are called *hotspot areas*. By covering hotspot areas, grids around Beijing Capital International Airport can also gain high density (with >$100\,vehicles/km^{2}$). Hotspots area and nodes density vary significantly during workdays compared with those during weekends. During weekends, nodes become more concentrated up to 400 $vehicles/km^{2}$.

Fig. 1 Average vehicles density for one day

Other than non-uniform distribution of vehicles density, connectivity among vehicles is another key factor of network clustering. Clustering coefficient is a metric of how well the neighbors of a node are connected and is therefore an effective criterion to measure clustering property. A clustering coefficient closer to 1 indicates better connectivity in the network. Suppose set of neighbors of node A at time t is denoted as N_A. E_{N_A} denotes the set of connections between all nodes pairs in N_A happened at t. The clustering coefficient is given by

$$COEFFICIENT_A^t = \frac{2\|E_{N_A}\|}{(\|N_A\|-1)\|N_A\|}$$

Figure 2 shows the topological distribution of the average clustering coefficient (denoted as C in the figure) during one day in the network. Specifically, clustering coefficient near the central urban area is higher than that in remote areas. Some areas reach a peak (about 0.4) , which indicates better connectivity. A closer look at the distribution shows that most of those areas have higher vehicles density and are the hotspot areas mentioned above. The average clustering coefficient in the whole network is relatively low because of road or obstacle constraints in urban DTNs. In most cases, a taxi can only communicate with adjacent vehicles on the same street, thereby reducing the chances of contact between neighbors of a specific vehicle. However, because of high aggregation in hotspot areas, vehicles can connect tightly in such areas. Thus, in large-scale urban DTNs, vehicles tend to strongly cluster in hotspot areas.

Fig. 2 Average clustering coefficient in Beijing's urban DTNs

Clustering based on *hotspot*s will become inefficient because the hotspots coverage area is quite large and unstable and its node density reaches $400\, vehicles / km^2$. In addition, location-based algorithms [23, 24] need to identify the hotspots in advance, which makes these algorithms inflexible. Dynamics of nodes and the network scale are an great challenge particularly in DTNs.

In this case, clustering based on mobile patterns[21] can be a better choice. As contact strength is related to aggregation degree, clustering nodes in logical groups according to contact strength will provide a direct expression of the urban clustering character and can adapt well to a large-scale dynamic network.

2.2 Contact Characteristics

ICT is the interval time between two continuous contacts of the same pair of nodes. ICT implies the contacts frequency of two nodes. Researchers focus on the ICT characteristics of different mobility patterns in DTNs, and numerous ICT distribution results of both theoretical and real-world mobility models are given. In addition, many ICT distribution models are created as exponential distributions. According to [17, 22, 25] , the ICT between two nodes generates exponential distributions, and the exponential parameter relates to the motion characteristics. Simulations in [17, 26]based on Radom Waypoint (RWP) mobility model and Radom Direction (RD) mobility model support the results. For the real-world mobility traces, the ICT between two vehicles is modeled in [27] and follows an exponential distribution.

Figure 4 plots CCDF of ICT in 2010.6.13 (with 86400 s) of Beijing urban taxi record. From the very beginning of the plotting period (~10000 s), the plot is almost a straight line with a negative slope on log-log scale. The tail of the distribution decreases rapidly after this range. Such a fact indicates that the ICT distribution in urban DTNs is an exponential nature. i.e. $P(X > t) = e^{-\lambda t}$. Only about 15% of ICTs last longer than 10000s in the plot, which means that most of contacts in the network can reappear after predictably short time.

Fig.3 CCDF of ICT in Beijing's urban DTNs

The contacts between vehicles in Shanghai were analyzed in [27, 28] and were found to havesimilar ICT nature. The contact probability formula based on ICT distribution is:

$$P(X > t) = e^{-\lambda t}$$

λ reflects the contact strength between nodes. Furthermore, a static λ was set for the whole network by fitting the global ICT distribution. The method measures the contact strength intuitively and efficiently, and has been frequently used.

In [22] , ICTs of node pairs was found to follow exponential distributions, too. Three basic assumptions on ICT theoretical analysis were given:

Assumption 1: The probability that any two mobile nodes will make contact two or more times in a small duration of time Δt is low.

Assumption 2: The probability that two mobile nodes will accomplish one contact within Δt is approximated as $\lambda \Delta t$.

Assumption 3: In the non-overlapping time durations, the contacts of any two mobile nodes are independent.

When the three assumptions are satisfied, the ICT of node pairs also follow exponential distributions with different exponential parameters, related to their contact times n and the sum of ICTs $\sum ICT_{ij}$. The exponential parameters for each pair of nodes respectively are set as follows:

$$p_{ij}(t) = P_{ij}(X \le t) = 1 - e^{-\frac{n}{\sum ICT_{ij}} t} \qquad (1)$$

We assume the ICT distribution follows a exponential nature, that is, the contact probability in time t is $P(X > t) = 1 - e^{-\lambda t}$. By fitting some node pair samples with the formula, the λ fit for a single pair of nodes varies significantly from 0.0005 to 0.005. Figure 4 shows the exponential distributions fitting for each pair of nodes. Thus, the parameters for each node pair according to their historical contacts, as shown in [hyt],may improve the accuracy of forecasting contact probability.

Fig. 4 ICT distribution of node pairs

For Beijing taxi trace data(2010.6.13), the average contact time is 44.6s and the average ICT is 2904.3 These results meet the three assumptions.

We use coefficient of determination R^2 to judge the similarity degree of theoretical curves and observed curves. R^2 is most often a number between 0 and 1.0, used to describe how well a regression line fits a set of data. An R^2 near 1.0 indicates that a regression line fits the data well, while an R^2 closer to 0 indicates that a regression line does not fit the data very well.

First, we fit the global dataset and obtain $\lambda = 0.00077527$. Then 500 pairs of nodes were chosen randomly, each will produce two theoretical curves and a actual curve and corresponding two R^2 . As shown in Figure 5, the average R^2 for setting the exponential parameter is 0.822, and that for the global fitting method is 0.717.

Fig. 5 Comparison between the R^2 of the two methods

Compared with the global fitting method [27]of one parameter fitting of the ICT exponential distribution of the whole network, parameter estimations for each pair of nodes[22] according to the dynamic network are more applicable and adaptive. Furthermore, when the scenario changes, global fitting parameter estimation need to be recalculated based on large historical data of the whole network. For the real-time parameter settings, the overhead of maintain clusters may increase, but distributed algorithm will balance the load.

Based on the analysis above, we propose the distributed algorithm which uses the ICT distributions for each node pairs.

3. CLUSTERING BASED ROUTING ALGORITHM

The proposed contact-predict clustering-based routing algorithm (CPCRA) is a contact probability based clustering routing algorithm. Contact probability is calculated by using the method in Section II. Thus, CPCRA only requires the local information (contact times and sum of ICTs) instead of overall information, and nodes with higher contact probability are organized into clusters. After clustering, the routing procedure is divided into intra- and inter-cluster routing. This algorithm is used to control network overhead and improve transmission efficiency.

In this section, we first introduce the local data structure of a node and relative symbols; and clustering criterion is given in Section 3.2. The distributed clustering procedure and the routine procedure are described in Sections 3.3 and 3.4, respectively.

3.1 Data Structure and Symbols

CPCRA is executed on every single node. Every node creates or updates its local information upon different events, e.g. *happen to contact*, *periodic update*, *record timeout*. As in Formulas 2, the local information includes node *id* and cluster identifier *cid*, *ContactTable* contains contact history information and *GatewayTable* contains the gateway information to other clusters from its own cluster.

$$\begin{cases} (id, cid, ContactTable, GatewayTable) \\ ContactTable = \{Contact\operatorname{Re}cord_1, ..., Contact\operatorname{Re}cord_n\} \\ GatewayTable = \{Gateway\operatorname{Re}cord_1, ..., Gateway\operatorname{Re}cord_n\} \\ Contact\operatorname{Re}cord = (id, cid, contactTimes, sumICT, contact\Pr ob) \\ Gateway\operatorname{Re}cord = (cid, id, timeout) \end{cases} \quad (2)$$

The detailed information mentioned above is shown in Table 1， these symbols may be used in the following paragraphs.

Tab. 1 Detailed information of the data structure

Symbol	**Description**
ID_i	Unique identifier of node i
p_{ik}	The contact probability of node i and node k in forecast time.
CID_i	Unique identifier of cluster which contains the node i
$ContactTable_i$	The data structure stores contact records of node i
$GatewayTable_i$	The data structure stores the gateway information from node i 's perspective.
$GID_i^{CID_c}$	Gateway id to cluster ^{CID_c} .The node in the same cluster of node i has the maximum probability to cluster ^{CID_c} .
$p_i^{CID_c}$	The Maximum probability of node i contacting with any node in cluster c .
$INFO_k$	The contact information of local node with node k , including contact times, sum of ICTs.
n_{ij}	Contact times between node i and node j
$\sum ICT_{ij}$	The sum of ICTs between node i and node j
η	The contact probability threshold to join a cluster.
$timeout$	A time mark to decide whether a gateway record is timeout.

3.2 Clustering Criterion

We propose a clustering criterion based on contact strength. The criterion differs from that of other clustering method considering historic factors, e.g., clustering by region, CPCRA uses $p_{ij}(t)$ (contact probability between nodes in a certain time t) as the clustering rules. Each node pair calculating the contact probability and directly reflects the mobility of nodes in DTNs in real time. $p_{ij}(t)$ is proved in [22] as Formulas 1.

$p_{ij}(t)$ is sometimes simplified as p_{ij} . n and $\sum ICT_{ij}$ will be updated when node i and j make contact in which then updates $p_{ij}(t)$ in real time.

$p_{ij}(t)$ can reflect the contact strength. We set the probability threshold η , and group nodes with a contact probability higher than η into clusters. A cluster can be represented as a collection of nodes:

$$C = \left\{ ID_k \middle| k = 1, 2, ..., n; \forall i, j \in \{1, 2, ..., n\}, p_{ij}(t) > \eta \right\}$$

According to the clustering rules, two parameters will affect clustering effect, namely, forecast time *t* and contact probability threshold . t and η interact with each other. When η is constant, a shorter t means a lower $p_{ij}(t)$ and fewer nodes will meet the clustering criterion. Smaller clusters will then be generated. However, the clusters will become more stable. As a consequence, the overhead of intra-cluster *routing* and clustering maintenance will decrease whereas the overhead of inter-cluster routing will increase. On the other hand, increasing η can influence the contact strength in cluster and guarantee the efficiency of intra-cluster routing, but will add clustering fragments at the same time. The overall overhead of clustering and routing into account and a reasonable t and η should be configured.

3.3 Distributed Clustering Procedure

Figure 6 shows the flowchart of the clustering procedure, including information exchange upon contact, periodic updating of $ContactTable$ and $GatewayTable$, and deleting related a gateway record when it times out.

Fig. 6 Clustering generic flowchart

Unlike the traditional MANET application, the assumption that the network topology remains static in the initial clustering phase does not apply to DTNs. Because nodes in DTNs move randomly over time, which varies the network topology over short period. Therefore, the clustering algorithm in DTNs should be distributed, heuristic and self-maintaining. The concept of cluster head is not introduced in the algorithm which functions as a coordinator in most clustering designs. To some degree, nodes are homogeneous because of their communication and computation abilities in DTNs. Determining which node qualifies as a cluster head without redundant bottleneck is difficult.

The detailed process will be introduced from the view of a single node i.

3.2.1 **Initialization**

Initially, each node generates an isolated cluster only, including the node itself and the cluster ID is set equal to the node ID. The node then initializes data structure, i.e. $ContactTable$ and $GatewayTable$.

3.2.2 **Happen to Contact**

When two nodes move into each other's communication range, they happen to contact. The node will execute the process shown in Figure 7. When node i is in contact with node j, node i will perform the following processes:

Fig. 7 Flowchart during contacting

First, the cluster Information of the other node is obtained. Upon contact with node j, node i will add or update the contact record with node j and re-calculate the contact probability.

The cluster ID is compared to determine whether they are in the same cluster.

If the two nodes belong to the same cluster, node i will update or insert the record in $GatewayTable_i$. For example, a record stores the information about the gateway node to cluster CID_m. Node i itself will decide whether or not node j can be a better gateway to cluster CID_m. First, node i identifies the maximum contact probability to cluster CID_m in $ContactTable_j$, which is denoted as $p(t)_j^{CID_m} = MAX\left\{p(t)_{jk} \middle| \forall k, CID_j^k = CID_m\right\}$ If $p(t)_{ij} \times p(t)_j^{CID_m} > p(t)_{iGID_i^{CID_m}} \times p_i^{CID_m}$, node j can be a better gateway for node i to cluster CID_m. As a result, node i will update the relevant record and reset the timeout timer. In case node i has no gateway to cluster CID_n, node j will be set as such a gateway if node j has any record which belongs to cluster CID_n in $ContactTable_j$.

If they do not belonged to the same cluster, node i will decide whether to join node j's cluster. The basic requirement is to meet the *clustering criterion* given in Section 3.2. In addition, node will join a more stable cluster. A higher minimum contact probability in a cluster indicates higher cluster stability. When node i joins the other cluster, it will set $CID_i = CID_j$ and $GatewayTable_i = GatewayTable_j$.

3.2.3 **Periodic Update Records**

In urban DTNs, the contact probability of a nodes pair varies with time. To obtain a more precise real-time measurement in real time, the algorithm will need to regularly update the records in $ContactTable$ and $GatewayTable$. A flowchart of the periodic record updating is shown in Figure 8.

Fig. 8 Flowchart for periodic record updating

First, the local node i updates p_{ij} in $ContactTable_i$ with the latest information By re-calculating the contact times n and $\sum ICT_{ij}$, the node can update in $INFO$ field.

Second, the node itself will go through clustering criterion check. If the node passes the criterion, it will remain in the original cluster it belonged to. Updating $ContactTable$ may change node i 's probability of routes to the other clusters. Therefore, node i will also update $GatewayTable_i$ to set the new or better gateway to other clusters. The gateway in $GatewayTable_i$ must be in the same cluster of node *i*.

If the node did not pass the *clustering criterion* check, it will leave its original cluster and construct a new isolated cluster which only includes itself. In such a situation, the node will keep its own $ContactTable$. $GatewayTable$ will be set empty, since the old gateway records are meaningless to the newly established cluster.

3.3 *Routing Procedures*

After clustering, each node stores enough information about its cluster members and gateways. Routing in that network no longer follows routing protocols which execute only one type of strategy in the whole network. Inside the cluster, because any two nodes have a high contact probability, a node will carry the packets until the desired destination appears in its communication range. Then it will directly deliver the packets in one hop. If the current node and the destination node do not belong to the same cluster, packets have to be relayed across more than one cluster for successful delivery. In that case, every cluster will be regarded as a more coarse-grained "node". A flooding-based algorithm will be executed among the clusters. More specifically, each cluster can only accept one copy of a packet. This condition ensures the delivery rate of messages and reduces redundant copies in the network.

3.3.1 **Intra-cluster Routing**

If the destination node has a record in the local node's $ContactTable$, and the two nodes are in the same cluster, intra-cluster routing algorithm will be executed. The local node directly delivers the packet to the destination node, which is a single-copy routing strategy in DTNs. Considering high contact probability inside a cluster, the local node can simply route decision making and ensure acceptable delay in the intra-cluster routing algorithm.

3.3.2 **Inter-cluster Routing**

If the local node and the destination belong to the different clusters, inter-cluster routing algorithm will be executed. The following two cases are considered.

In the first case, the destination node j has a record in local node i 's $ContactTable$, but $CID_i \neq CID_j$, node i further explores its $GatewayTable$. If the node finds a gateway $GID_i^{CID_j}$ to cluster CID_j, it will regard the gateway as next hop and execute intra-cluster routing between them. After the packet is received, the gateway will relay the packet to any node which belongs to cluster CID_j.

In another case, node i cannot find any record of node j in $ContactTable_i$, or it discovers that they belong to different clusters but cannot pinpoint the gateway to cluster CID_j. At this moment, flooding-based algorithm will be used for routing. First, node i marks the relayed packet with the type of routing strategy. Upon contact with any other node k, if $CID_i \neq CID_k$, and node i finds that no copy of the packet in cluster CID_k is found, node i will relay the packet to node k. Then node i will record that cluster CID_k received one copy. As a result, if node i contacts with a node from cluster CID_k, it will not relay the copy again in the future.

4. SIMULATION

In this section, we design and execute three simulation experiments to evaluate the proposed algorithm. First, we evaluate the clustering performance and determine the parameter settings (t, η). Second, we compare delivery ratio and overhead of traditional single-strategy routing protocols with the proposed algorithm in the large-scale urban DTN scenario to demonstrate its routing performance. Third, by setting up scenarios with different network scales, we reveal the scalability of the algorithm compared with other protocols. The protocols with which we compare the CPCRA and the reason for choosing them are shown in Table 2. All simulation scenarios are extracted from the data set we describe in Section 3, which is one of the largest urban DTNs data sets in the world.

Tab. 2 Single-strategy protocols used in simulation

Protocol	Reason for comparison
Direct Delivery	We refine the same strategy in intra-cluster routing from this protocol
Epidemic	Also using flooding strategy as our inter-cluster routing
PRoPHET	A two-hop forwarding strategy are involved based on contact criterion

4.1 Clustering Evaluation

In this section, we try to evaluate the clustering performance. The average scale of clusters, or the average node numbers in a cluster, is chosen to evaluate the clustering. The average scale of clusters is affected by clustering rules and is an important criterion for evaluating clustering performance. Whether or when the cluster becoming stable is another criterion for evaluation.

Quantifying a certain clustering criterion in different scenarios cannot be done. For example, in a small scale network with high node density, small-scale cluster may be a better message delivery method. [29] proposes a method of comparing the clustering performance when change its own factor. We can evaluate the clustering algorithm by this approach.

Simulation parameters are shown in Table 3, in which the variable parameters are node number, contact probability threshold and contact forecast time.

Tab. 3 parameter settings of clustering evaluation

Parameter	**Value**
Simulation time	40000s
Buffer Size	2.5Mb
Message size	50Kb-100Kb
Transmission range	200m
Inter-message creation time	5s
bandwidth	256K
Node number	500-2000
area	20000m*20000m
η	0.15, 0.3
t	200s, 400s

The average scale of clusters reflects the granularity of clustering, and the curve of the change of the average scale of clusters can indicate whether a cluster can become stable. Two node scales, namely, 1000 and 1500, are chosen to compare the clustering performance by changing a single parameter, contact probability threshold or contact forecast time. Figures 9 and Figure 10 illustrate the change of the average scale of the cluster with time.

Figure 9 illustrates the curve of the cluster scale variation by changing η from 0.3 to 0.15, at a forecast time of 400 s. A larger network scale, i.e. node number indicates that the average scale of clusters becomes larger. High node density increase the opportunity to contact so that the contact probability between nodes becomes higher than η, which also increases the degree of convergence. The η influences the average scale of clusters directly. In Figure 5 (left), the peak value of the scale is 13.6 with $\eta = 0.15$, while the peak value is 9.4 with $\eta = 0.3$

Fig. 9 Average scale of clusters with different threshold(left: 1000 nodes; right: 1500 nodes).

Fig. 10 Average scale of clusters in different forecast time(left: 1000 nodes; right: 1500 nodes).

Figure 10 uses the same contact threshold $(\eta = 0.3)$, but its forecast time t changes from 400s to 200s. In Figure 10 (left), when $t = 200$, the cluster scale peak value is 6.8, and when $t = 400$, the peak reaches 9.4. This result is due to the fact that forecast time increase the contact probability value, which cause more nodes to satisfy the clustering criteria and join clusters.

In every case, the curve can achieve stability after 2000 s. It shows that the clustering algorithm can cluster the network effectively and stably.

Based on the analysis above, the proposed clustering algorithm can organize the network into clusters which have certain degree of granularity and achieve stability. Then, the high contact probability of nodes in each cluster clarifies the task of each node. Thus, this clustering algorithm can cluster effectively and ensure the stability of the network topology.

4.2 Delivery Ratio and Overhead

In this simulation, we randomly select traces of 4500 vehicles from the central area ($20000m \times 20000m$) of Beijing. As revealed in Section 3, the central area of the city has numerous hotspot areas, which we believe is an appropriate scenario for the clustering based-routing algorithm. We investigate delivery ratio and overhead of different protocols under different buffer sizes. Detail parameter settings are shown in Table 4.

Figure 11 shows that most protocols have higher delivery ratio with increased buffer size. The proposed routing algorithm yields a higher delivery ratio than the other single-strategy protocols because in such a large-scale scenario, clustering strategy groups nodes into different clusters, which makes the network more coarse-grained. As a result, from a node's point of view, the network scale becomes much smaller. Furthermore, a node can make its forwarding decision inside or outside of the cluster more wisely which not only saves the buffer of local nodes, but also tends to the route with higher contact probability.

Tab. 4 Parameter setting in Simulation One

Simulation time	10000s
Simulation area	20000m*20000m
Number of nodes	3000
Communication range	200m
Packet size	Random(50k, 100k)
Buffer size	[0.5M, …, 5M]
Protocols	CPCRA($t = 400, \eta = 0.3$); Epidemic; Prophet; Direct Delivery

Fig. 11 Delivery ratio of protocols with different buffer sizes.

Overheads are shown in Figure 12, which indicates the average number of copies generated by the network until a packet is successfully delivered. Direct Delivery is a single copy protocol which dose not produce additional copy of packets. Epidemic Protocol duplicates a copy whenever a node contact with others, thereby producing the highest overhead among the protocols. In the clustering-based routing algorithm, nodes generate copies only when they meet the other nodes in different clusters where no copy of the specific packet exists, thereby controlling the overhead. On the other hand, the overhead of PROpHET Protocol is two times greater than than that of the proposed algorithm because of its high forwarding probability and one-hop decision making. As a result, the clustering-based routing algorithm yields a really low overhead between that of PROpHET and Direct Delivery Protocol, and has considerable performance in large-scale real-world application.

Fig. 12 Overhead of protocols with different buffer sizes

4.3 Scalability

Scalability is one of the major consideration of the routing protocol in large-scale DTN application. The network scale is mainly changed by the number of nodes in the network. A greater number of nodes involved in the network indicates greater overhead because many additional copies of packets and messages to synchronous routing information are generated. In this case, the network may encounter a bottleneck and the performance of most protocols may stagnate or deteriorate. In this simulation, the scalabilities of protocols are compared by varying the number of nodes. Detailed parameters settings are shown in Table 4.

Fig. 13 Delivery ratio of protocols with different network scales

Figure 13 shows the delivery ratio in the network which has different numbers of nodes. The performances of the single-strategy routing protocols do not improve as the number of nodes increases. In general, increasing nodes density can bring more contact opportunities which should lead to higher delivery ratio. However, if the network scale is large enough, the buffer capacity becomes limited during single-strategy routing because of their overall forwarding behaviors, which limit the increase of delivery ratio. The clustering-based routing algorithm can adapt to the increasing scale. If the clustering parameters are constant ($t = 400s, \eta = 0.3$), the number of clusters does not change considerably when the network scale increases. Instead, the size of each cluster, which is denoted by number of nodes inside each cluster, increases. Therefore, from a node's point of view, the buffer capacities of clusters increase. When executing inter-cluster routing, each cluster is considered a bigger "node". Thus, the network maintains the small number of "nodes" which have a higher buffer capacity, thereby improving routing performance.

Fig. 14 Overhead of protocols with different network scales

Figure 14 reveals the superior scalability of the proposed algorithm based on the other metric, namely, overhead growth rate. When the network is sparse, overhead of single-strategy protocols is close to that of clustering-based routing. However, as the number of nodes increases, overheads of those protocols, expect that of Direct Delivery Protocol, become dramatically high. On the other hand, the proposed algorithm can restrict its overhead growth rate. The reason for this condition is similar to the one we described above. In large-scale urban DTNs, the proposed algorithm can help reduce network overhead and achieve scalability thereby outperforming traditional single-strategy routing protocols.

5. CONCLUSIONS

In this paper, we analyzed the network characteristics of large-scale urban DTNs from Beijing. Nodes tend to cluster into hotspot areas and distribution of ICT exhibits an exponential nature in such kind of networks. To improve the scalability of routing protocols used in large-scale application, we proposed a clustering-based routing algorithm by using contact probability as the clustering criterion. We further derived the expression of contact probability based on analytical study of ICT. The simulation results demonstrated that the performance and scalability of the proposed algorithm exceed those of traditional single-strategy protocols, which means the proposed algorithm can be applied in large-scale urban DTNs.

Many aspects of this work can be explored in the future. First, the algorithm can be evaluated from the aspect of clustering in terms of its stability and maintenance overhead. Second, the algorithm can be improved to be a more adaptive by studying the influence of parameters. Moreover, our algorithm can be deployed in real-world scenarios to evaluate its effectiveness and suitability for application.

ACKNOWLEDGMENTS

This research was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 61170295, the Project of National Ministries Foundation of China under Grant No. A212010006, the Project of Aeronautical Science Foundation of China under Grant No. 2011AC51024, the Co-Funding Project of Beijing Municipal Education Commission under Grant No. JD100060630 , the Fundamental Research Funds for the Central Universities No. YWF-12-LXGY-001 ，and Network information and computing center of Beihang University.

References:

[1] K. Fall, "A Delay-Tolerant Network Architecture for Challenged Internets," in *SIGCOMM'03* Karlsruhe, Germany, 2003, pp. 27-34.

[2] K. Fall and S. Farrell, "DTN: an architectural retrospective," *Selected Areas in Communications, IEEE Journal on,* vol. 26, pp. 828-836, 2008-01-01 2008.

[3] S. Ahmed and S. K. Salil, "Characterization of a large-scale Delay Tolerant Network," in *Local Computer Networks (LCN), 2010 IEEE 35th Conference on*, 2010, pp. 56-63.

[4] S. Saha, N. Ganguly and A. Mukherjee, "Information Dissemination Dynamics in Delay Tolerant Network: A Bipartite Network Approach," *Proceedings of the third ACM international workshop on Mobile Opportunistic Networks,* 2012-01-01 2012.

[5] A. J. Kleywegt, V. S. Nori and M. W. P. Savelsbergh, "The stochastic inventory routing problem with direct deliveries," *Transportation Science,* vol. 36, pp. 94--118, 2002.

[6] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine, "Maxprop: Routing for vehicle-based disruption-tolerant networks," in *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, 2006, pp. 1--11.

[7] A. Vahdat and D. Becker, "Epidemic routing for partially connected ad hoc networks," *Technical Report CS-200006, Duke University,* 2000-01-01 2000.

[8] R. Agarwal, D. Motwani and Others, "Survey of clustering algorithms for MANET," *arXiv preprint arXiv:0912.2303*, 2009.

[9] C. Liu, Y. Liu, X. Ma, and J. Gao, "An Application Scheme of Publish/Subscribe System over Clustering Mobile Ad Hoc Networks," in *Wireless Communications Networking and Mobile Computing (WiCOM), 2010 6th International Conference on*, 2010, pp. 1--4.

[10] H. Wang, L. Mi, C. Xia, L. Lv, and M. Chen, "TLDV: Tree-Like Locator Distance Vector routing protocol for MANETs," in *Educational and Information Technology (ICEIT), 2010 International Conference on*, 2010, pp. V2--291.

[11] J. Whitbeck and V. Conan, "HYMAD: Hybrid DTN-MANET routing for dense and highly dynamic wireless networks," *Computer Communications,* vol. 33, pp. 1483--1492, 2010.

[12] C. Liu and J. Wu, "Scalable Routing in Delay Tolerant Networks," *MOBIHOC'07: PROCEEDINGS OF THE EIGHTH ACM INTERNATIONAL SYMPOSIUM ON MOBILE AD HOC NETWORKING AND COMPUTING,* pp. 51-60, 2007.

[13] S. Ahmed and S. S. Kanhere, "Cluster-based forwarding in delay tolerant Public Transport Networks," in *Local Computer Networks, 2007. LCN 2007. 32nd IEEE Conference on*, Dublin, Ireland, 2007, pp. 625 - 632.

[14] W. Gao and G. Cao, "On exploiting transient contact patterns for data forwarding in delay tolerant networks," in *Network Protocols (ICNP), 2010 18th IEEE International Conference on*, 2010, pp. 193--202.

[15] J. Whitbeck and V. Conan, "HYMAD: Hybrid DTN-MANET routing for dense and highly dynamic wireless networks," *Computer Communications,* vol. 33, pp. 1483--1492, 2010.

[16] I. Carreras, D. Miorandi and I. Chlamtac, "A simple model of contact patterns in delay-tolerant networks," *WIRELESS NETWORKS,* vol. 16, pp. 851-862, 2010.

[17] R. Groenevelt, P. Nain and G. Koole, "The message delay in mobile ad hoc networks," *Performance Evaluation,* vol. 62, pp. 210--228, 2005.

[18] Y. Li, P. Hui, D. Jin, L. Su, and L. Zeng, "Evaluating the impact of social selfishness on the epidemic routing in delay tolerant networks," *Communications Letters, IEEE,* vol. 14, pp. 1026--1028, 2010.

[19] L. Yong, J. Yurong, J. Depeng, S. Li, Z. Lieguang, and W. Dapeng, "Energy-Efficient Optimal Opportunistic Forwarding for Delay-Tolerant Networks," *Vehicular Technology, IEEE Transactions on,* vol. 59, pp. 4500-4512, 2010-01-01 2010.

[20] T. Hossmann, T. Spyropoulos and F. Legendre, "Know thy neighbor: Towards optimal mapping of contacts to social graphs for DTN routing," in *Proceedings - IEEE INFOCOM* San Diego, CA, United states, 2010.

[21] D. Ha and W. Hongyi, "Clustering and cluster-based routing protocol for delay-tolerant mobile networks," *IEEE Transactions on Wireless Communications,* vol. 9, pp. 1874-1881, 2010-01-01 2010.

[22] Y. Hu, H. Wang, C. Xia, W. Li, and Y. Yang, "On the Distribution of Inter Contact Time for DTNs," in *IEEE，LCN （Local Computer Networks）*, 2012.

[23] M. Kim, D. Kotz and S. Kim, "Extracting a mobility model from real user traces," in *Proc. IEEE Infocom*, 2006, pp. 1--13.

[24] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *Mobile Computing, IEEE Transactions on,* vol. 6, pp. 606--620, 2007.

[25] Y. Jiang, Y. Li, L. Zhou, D. Jin, L. Su, and L. Zeng, "Optimal opportunistic forwarding with energy constraint for DTN," in *INFOCOM IEEE Conference on Computer Communications Workshops, 2010*, 2010, pp. 1--2.

[26] X. Zhang, H. Zhang and Y. Gu, "Impact of source counter on DTN routing control under resource constraints," in *2nd International Workshop on Mobile Opportunistic Networking, MobiOpp 2010*, Pisa, Italy, 2010, pp. 41-50.

[27] H. Zhu, M. Li, L. Fu, G. Xue, Y. Zhu, and L. M. Ni, "Impact of Traffic Influxes: Revealing Exponential Intercontact Time in Urban VANETs," *IEEE Trans. Parallel Distrib. Syst*, vol. 22, pp. 1258--1266, 2011.

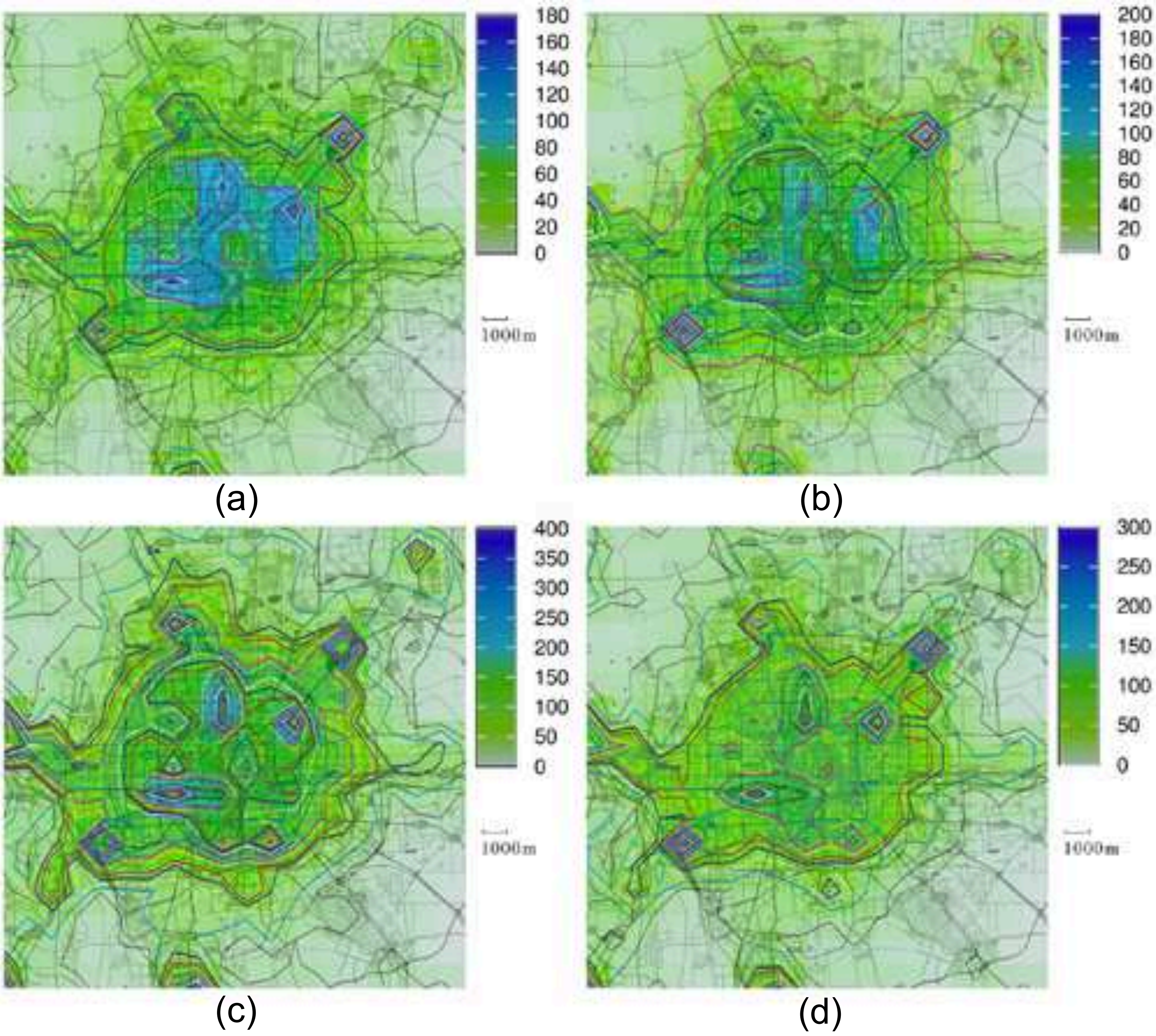
[28] X. L. Zhang, J. Kurose, B. N. Levine, D. Towsley, and H. G. Zhang, "Study of a Bus-based Disruption-Tolerant Network: Mobility Modeling and Impact on Routing," *MOBICOM'07: PROCEEDINGS OF THE THIRTEENTH ACM INTERNATIONAL CONFERENCE ON MOBILE COMPUTING AND NETWORKING,* pp. 195-206, 2007.

[29] A. B. McDonald and T. F. Znati, "A mobility-based framework for adaptive clustering in wireless ad hoc networks," *Selected Areas in Communications, IEEE Journal on*, vol. 17, pp. 1466-1487, 1999-01-01 1999.

Highlights (for review)

- We analyze the distributions of active vehicles in Beijing.
- Assume inter-contact time (ICT) follows an exponential distribution, two model method to estimate exponential parameter are compared, and one of which is chosen forecast contact probability.
- We proposed a contact-predict clustering-based routing algorithm(CPCRA) in Large-scale Urban DTNs and evaluate the clustering and routing performance.

Fig. 1 Average vehicles density for one day



(a)

(b)

(c)

(d)

Fig. 2 Average clustering coefficient in Beijing's urban DTNs

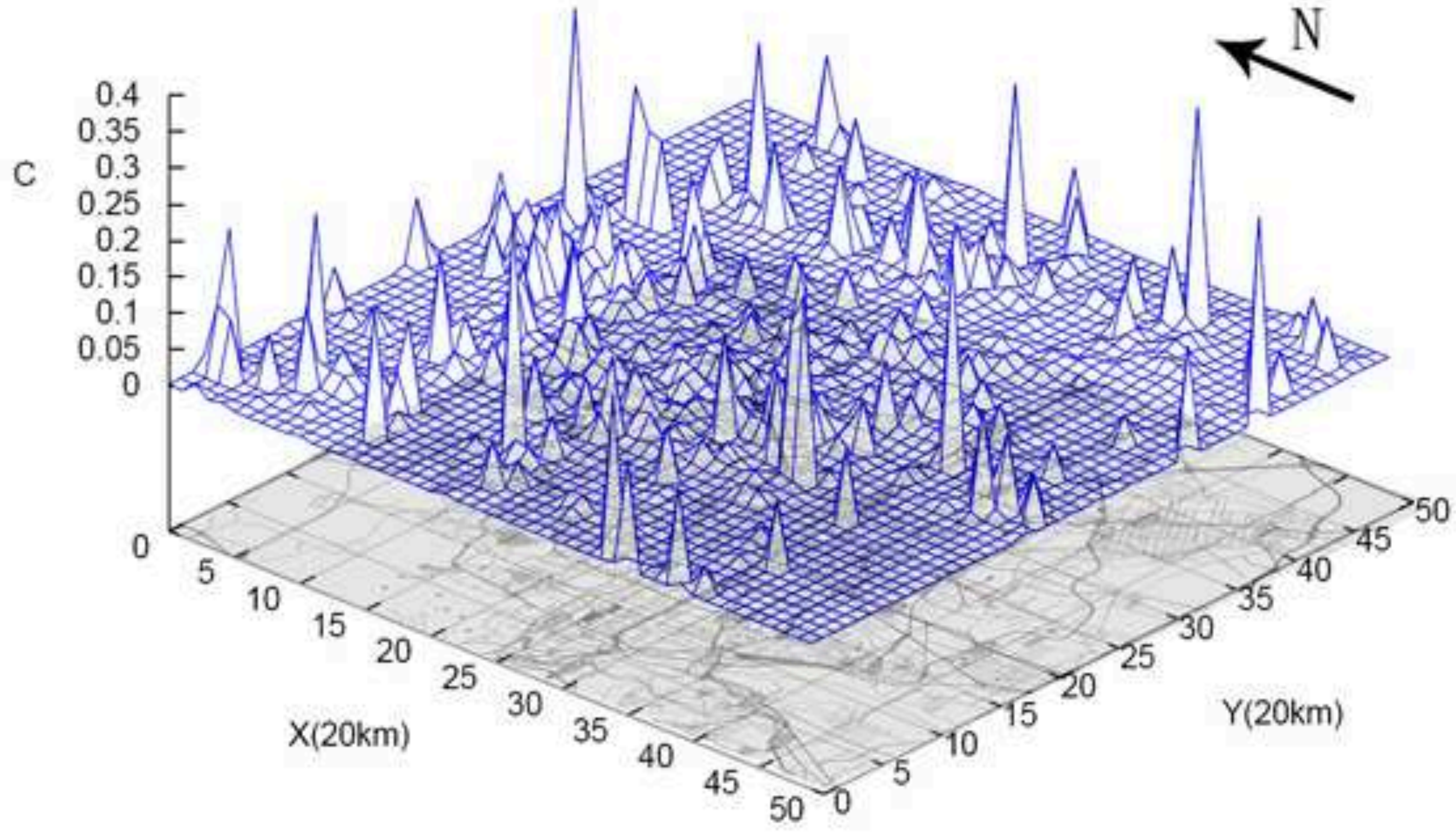


Fig.3 CCDF of ICT in Beijing's urban DTNs

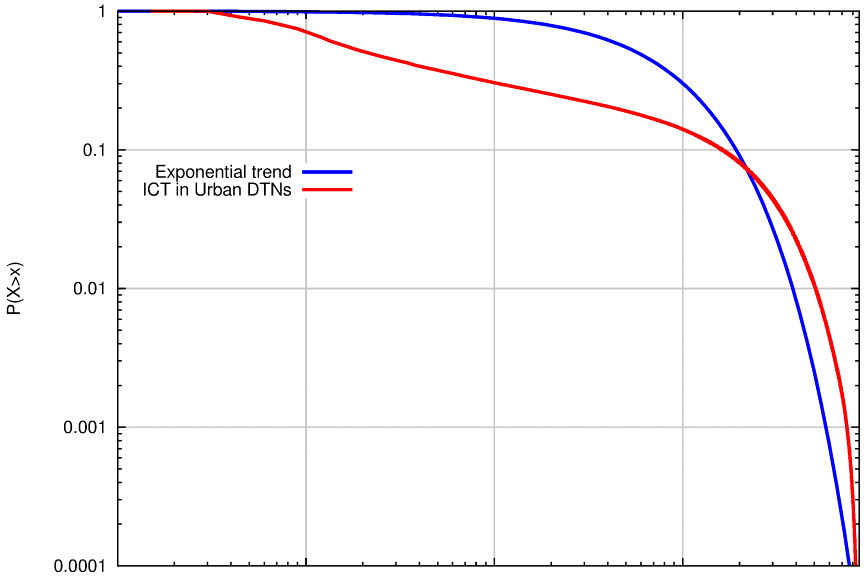


Fig. 4 ICT distribution of node pairs

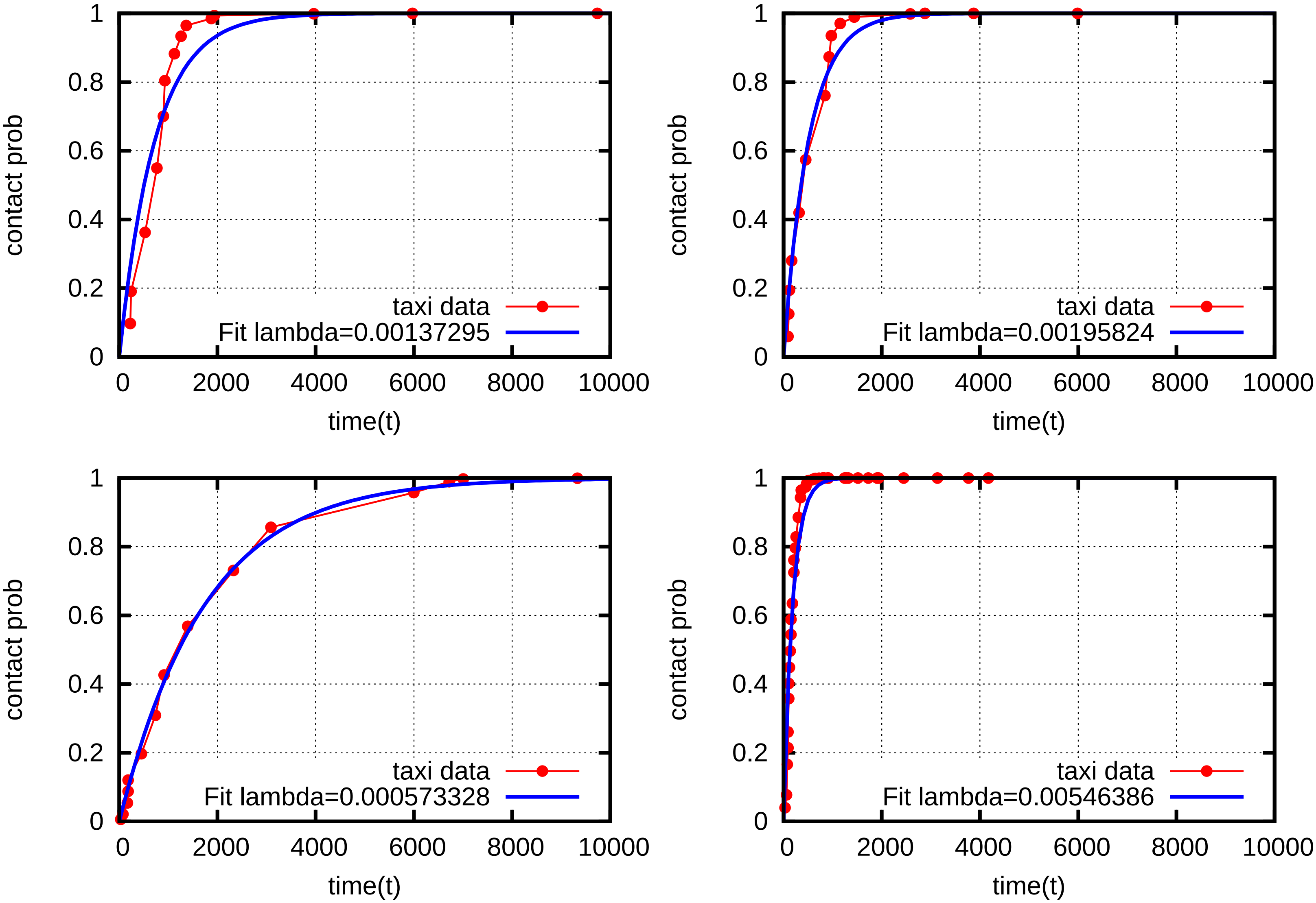


Fig. 5 Comparison between the R^2 of the two methods

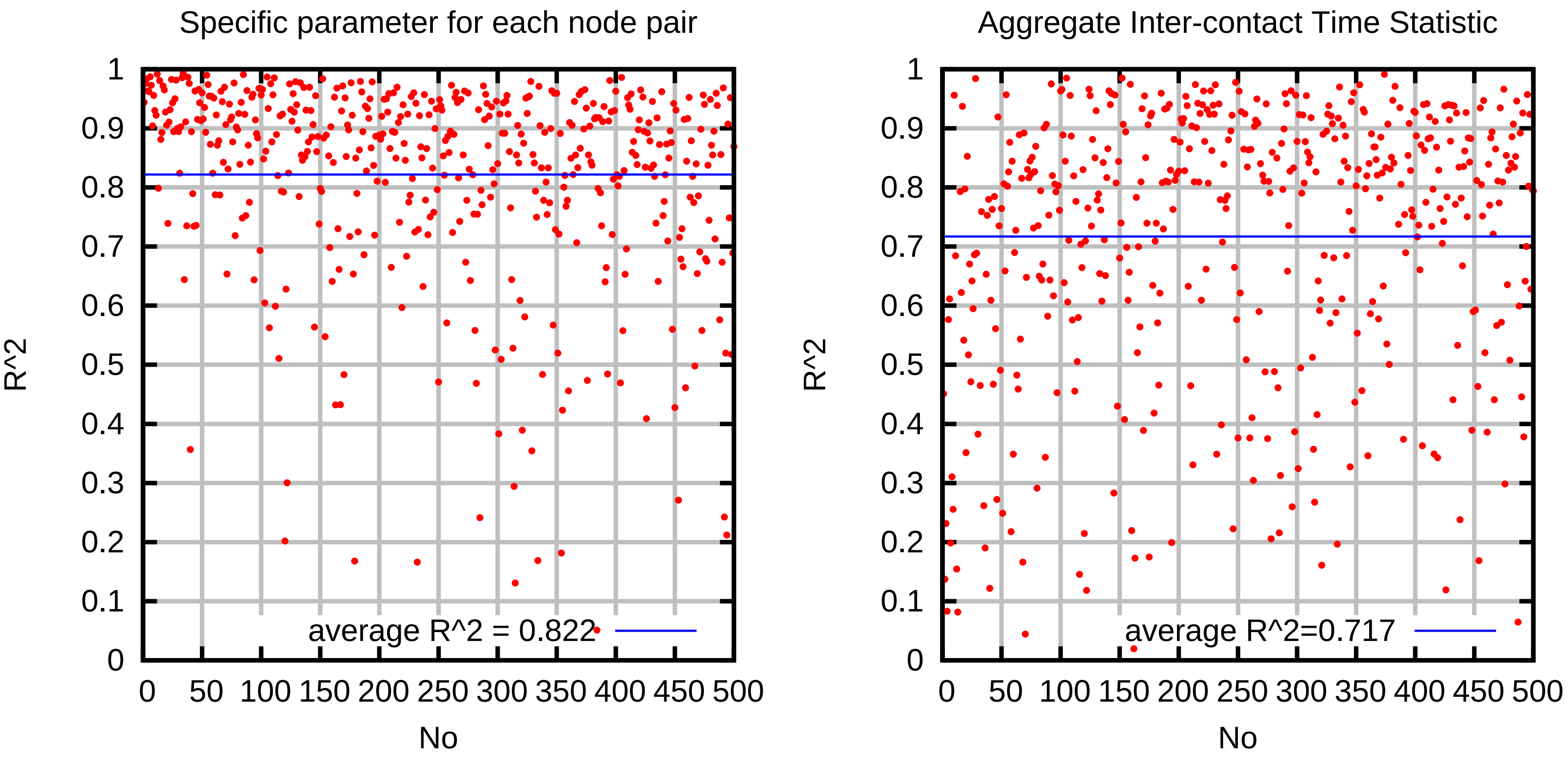


Fig. 6 Clustering generic flow chart

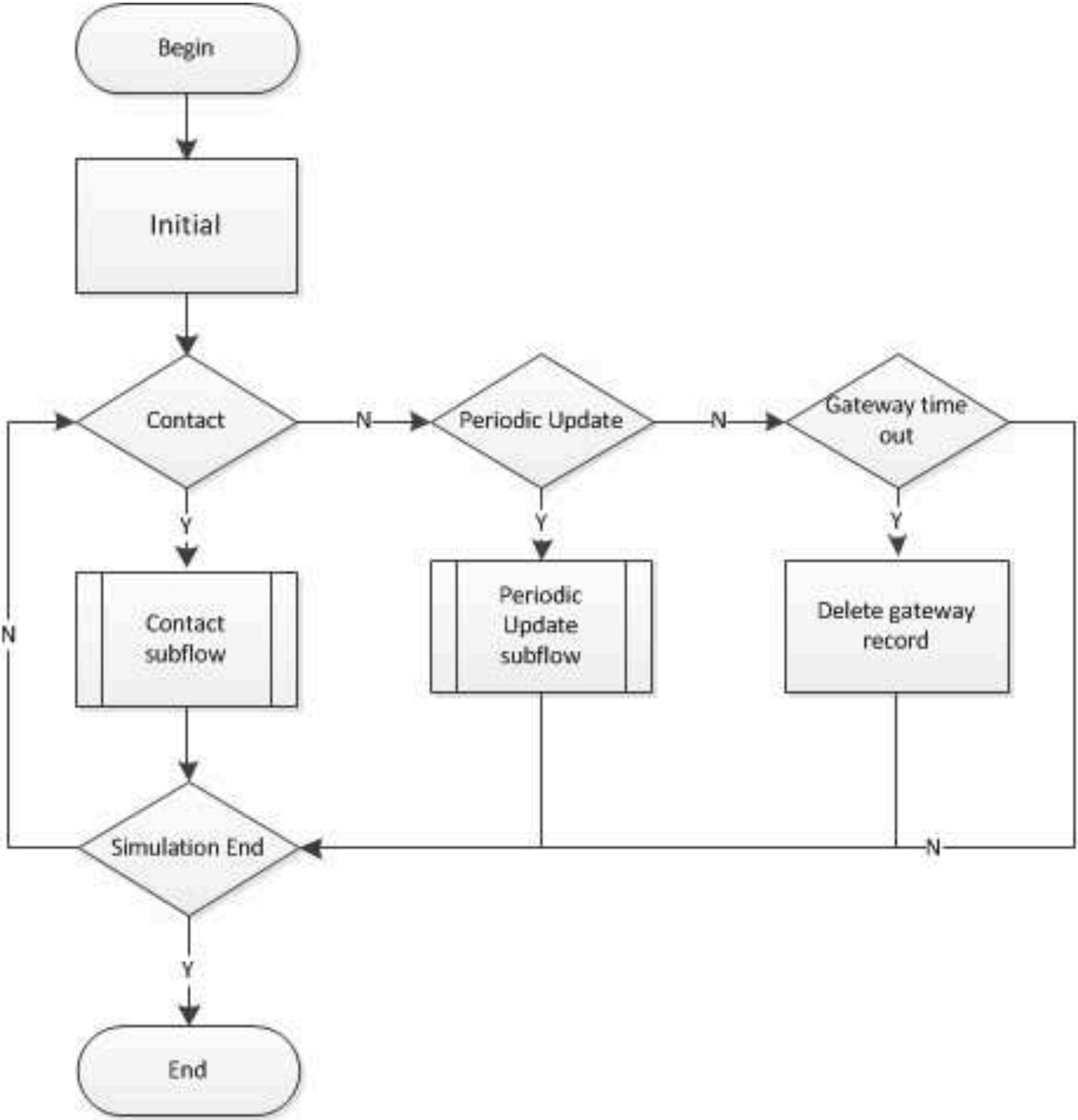


Fig. 7 Flowchart during contacting

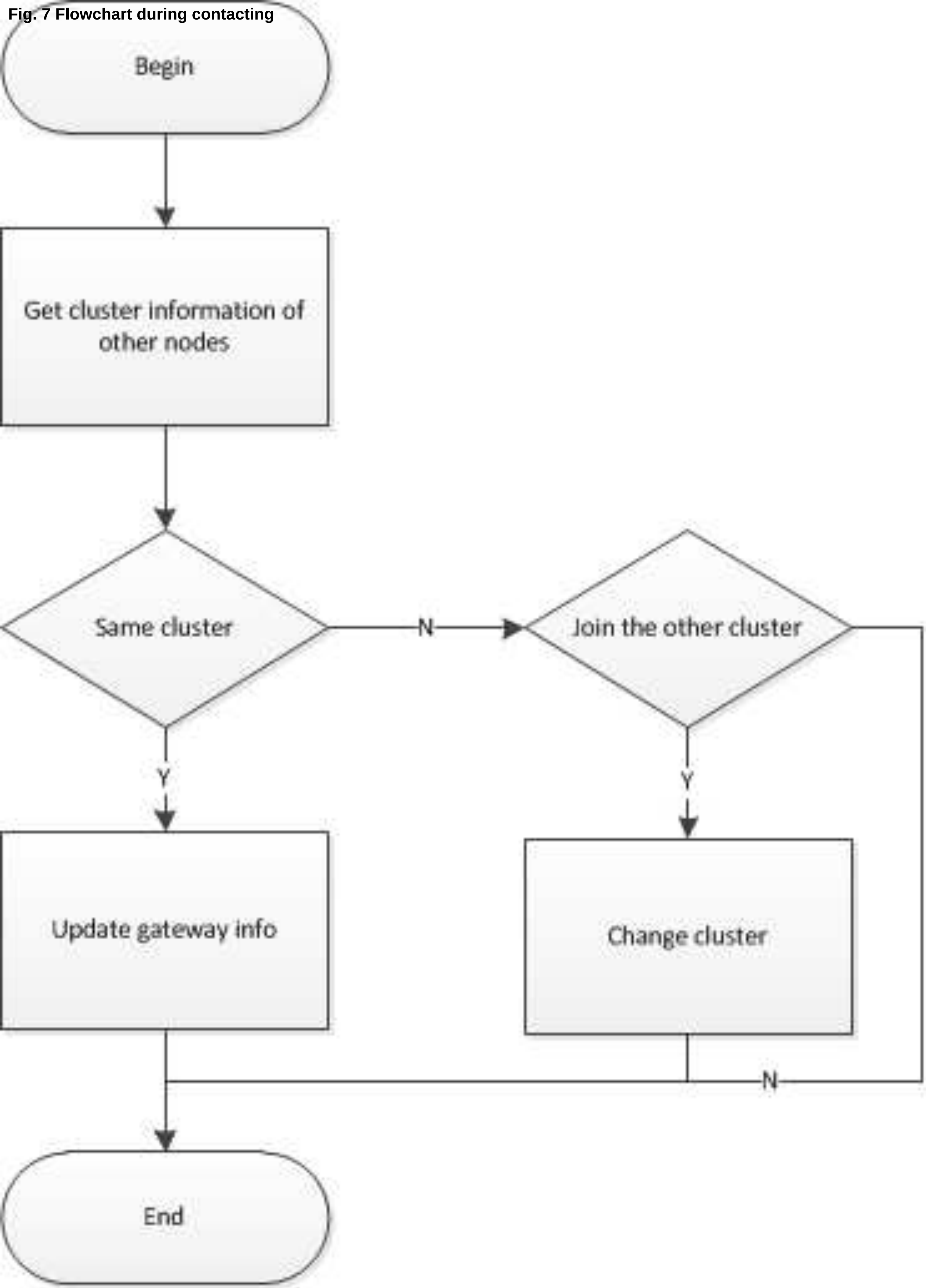


Fig.8 Flowchart for periodic record updating

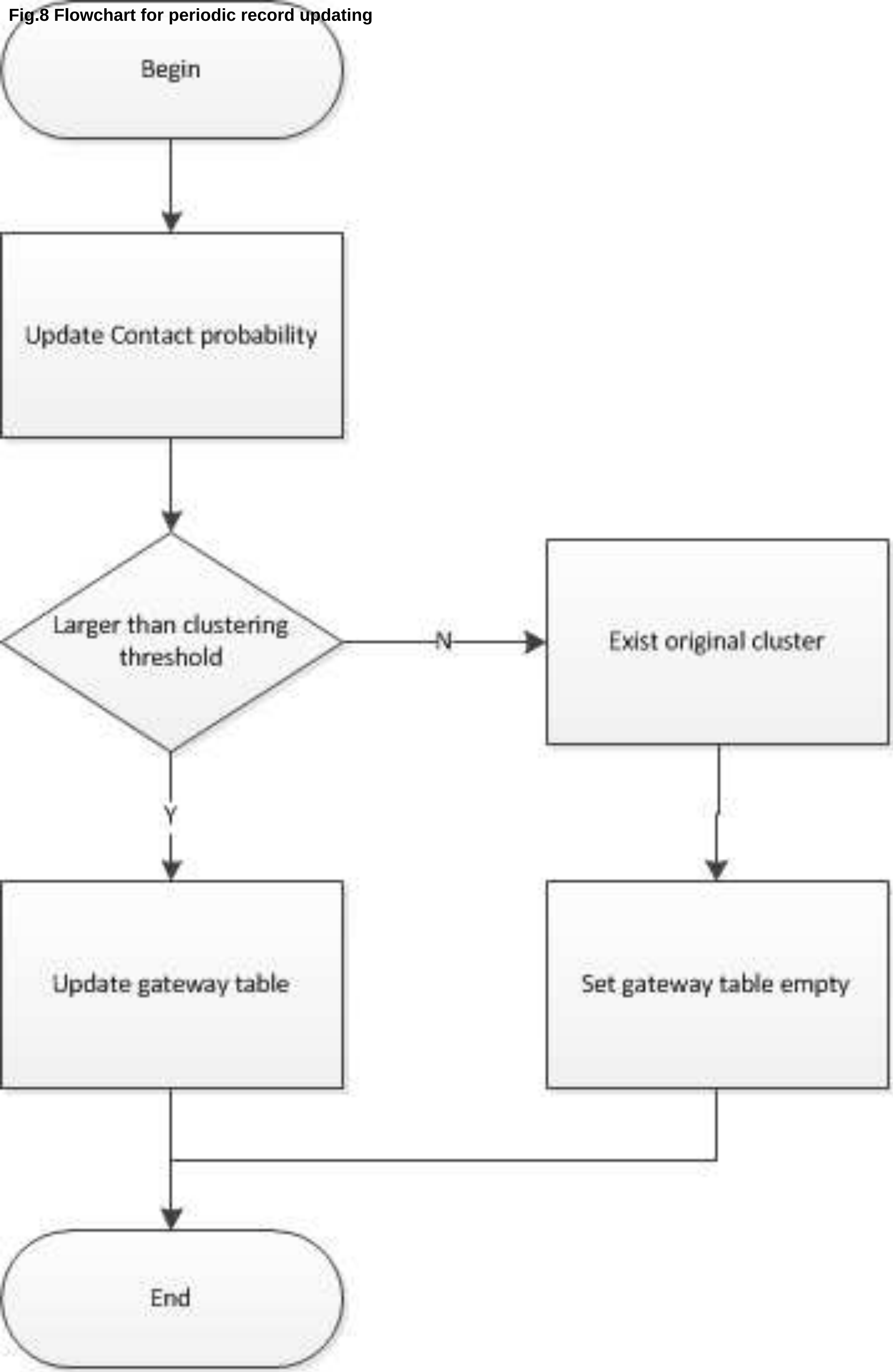


Fig. 9 Average scale of clusters with different threshold(left:

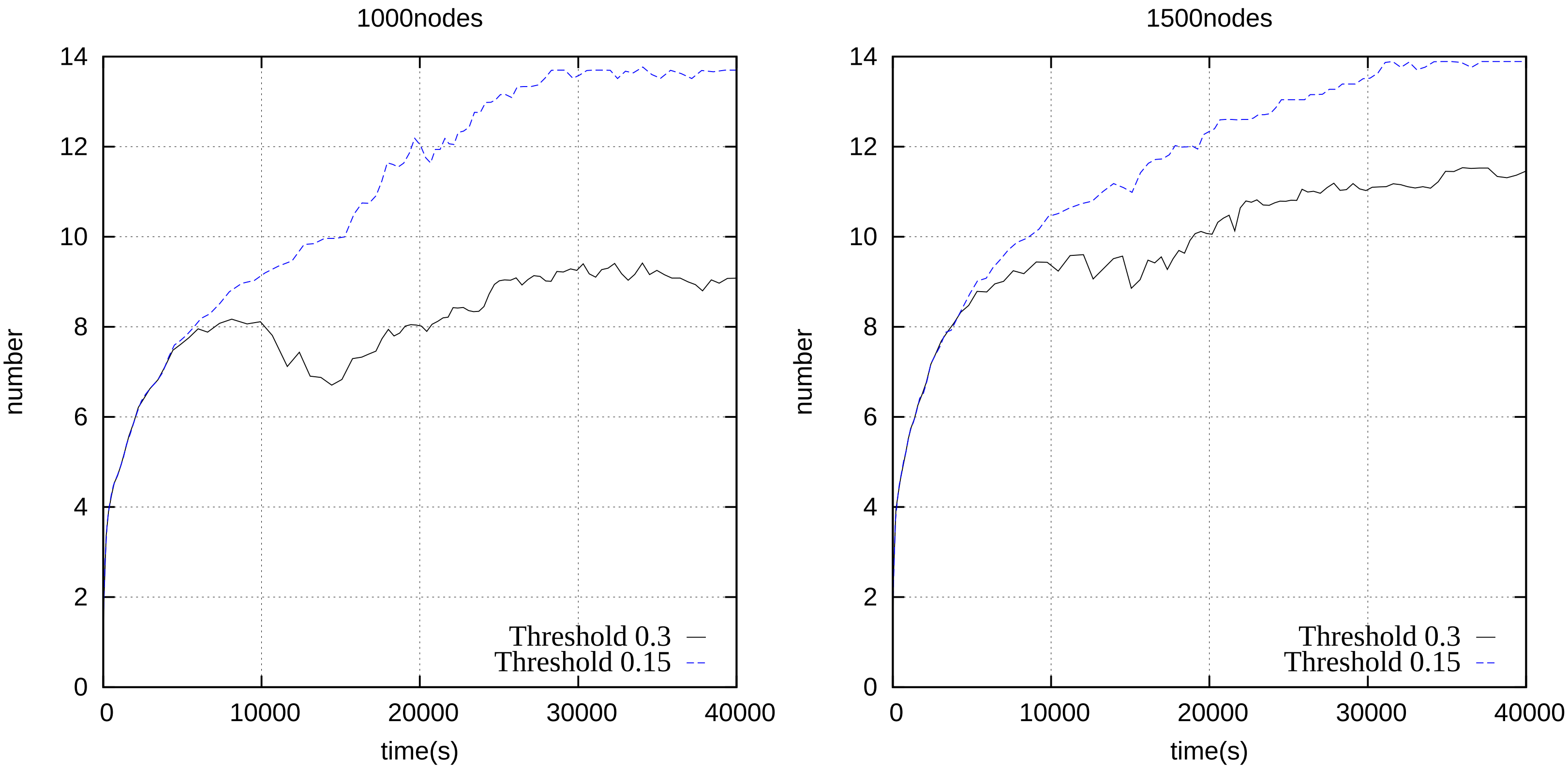


Fig. 10 Average scale of clusters in different forecast time(lef

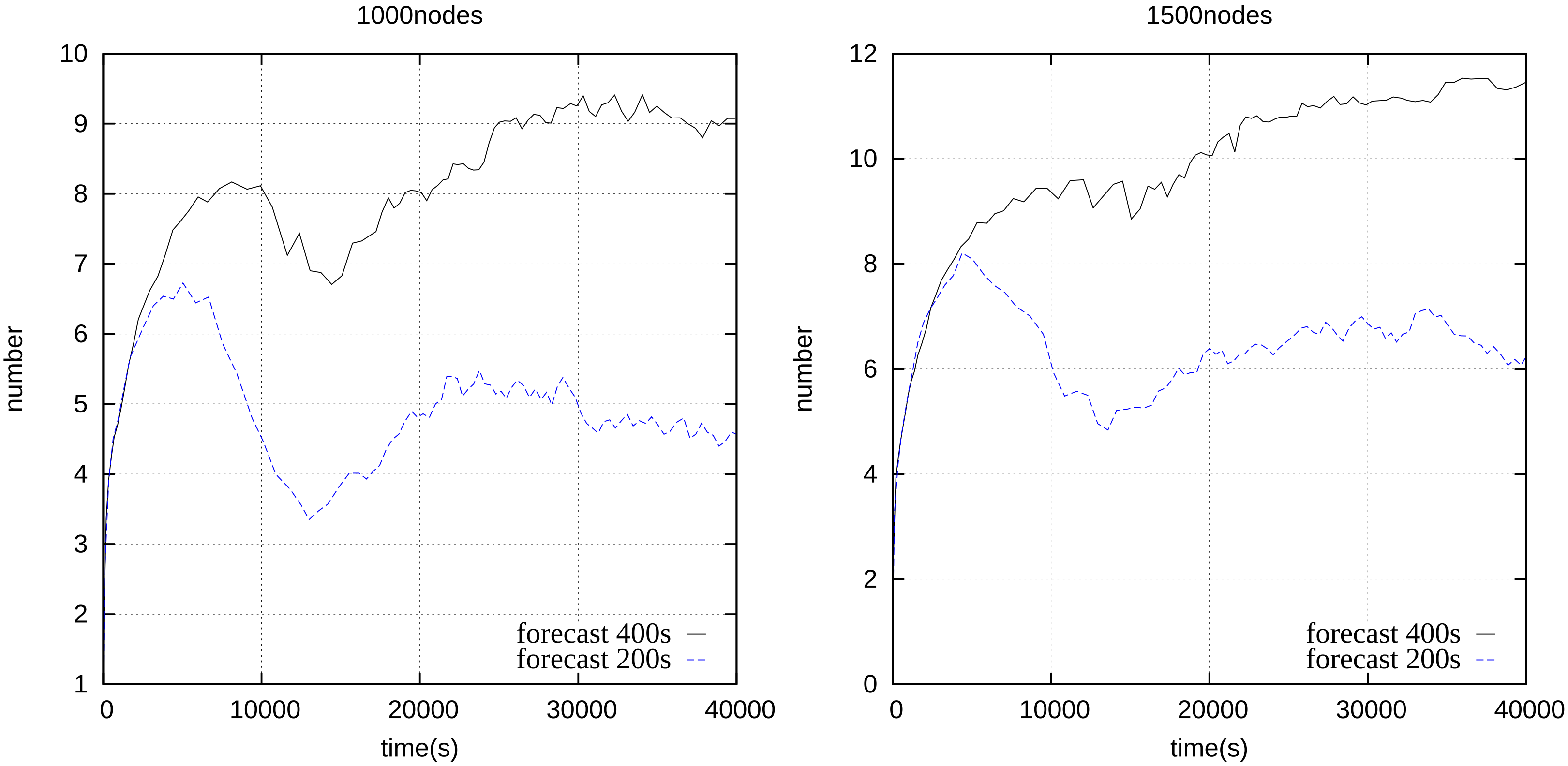


Fig. 11 Delivery ratio of protocols with different buffer sizes.

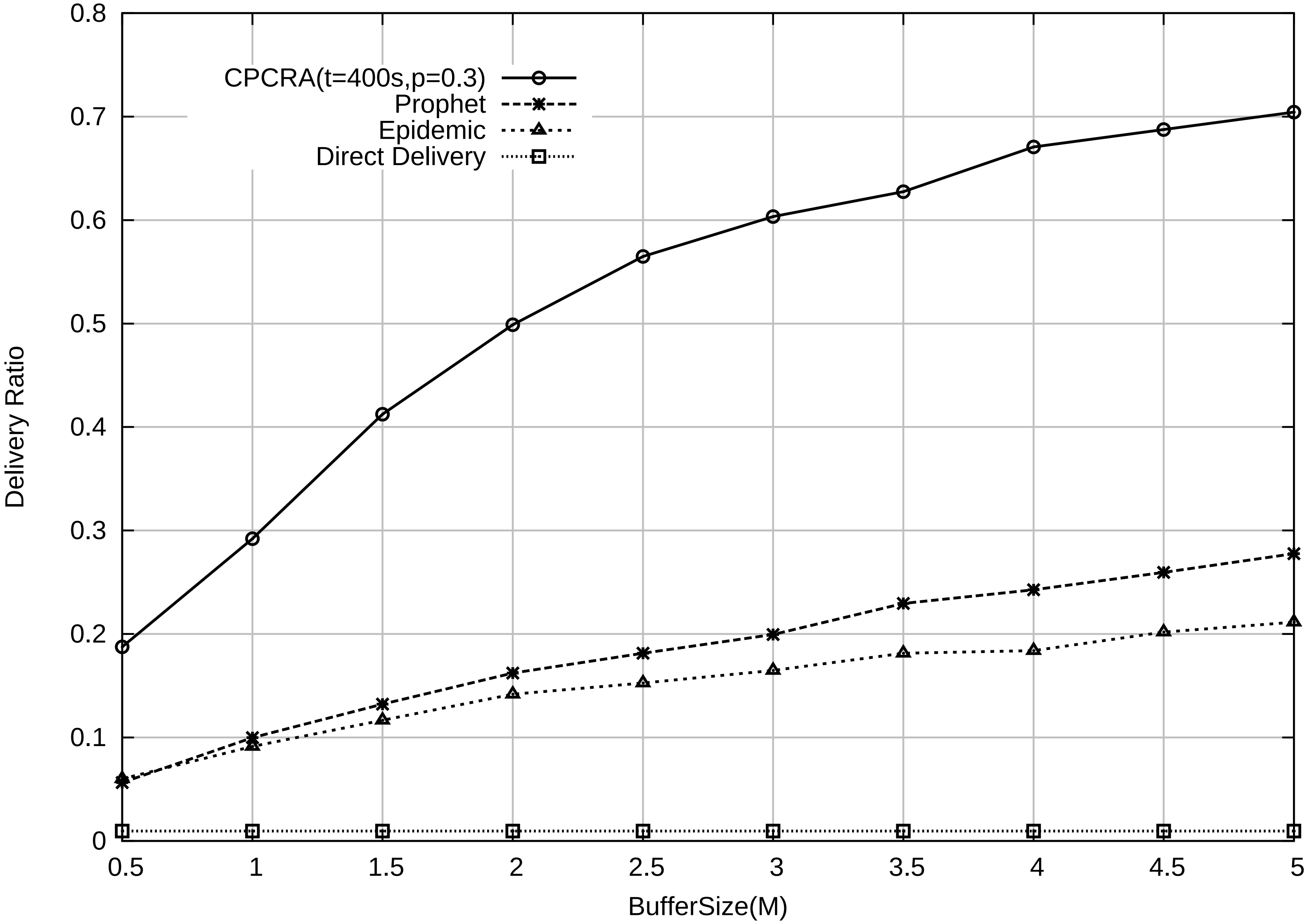


Fig. 12 Overhead of protocols with different buffer sizes.

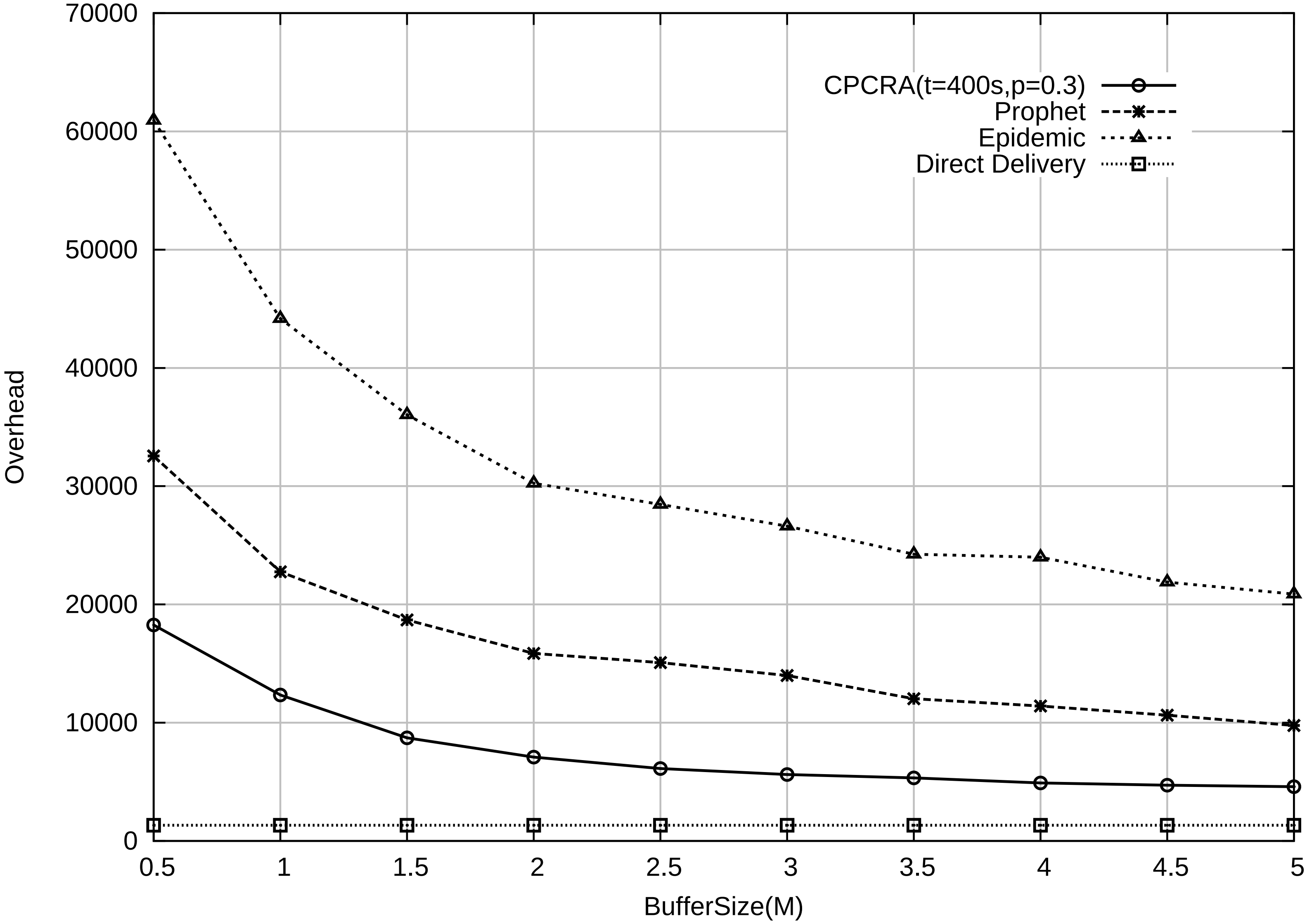


Fig. 13 Delivery ratio of protocols with different network scale

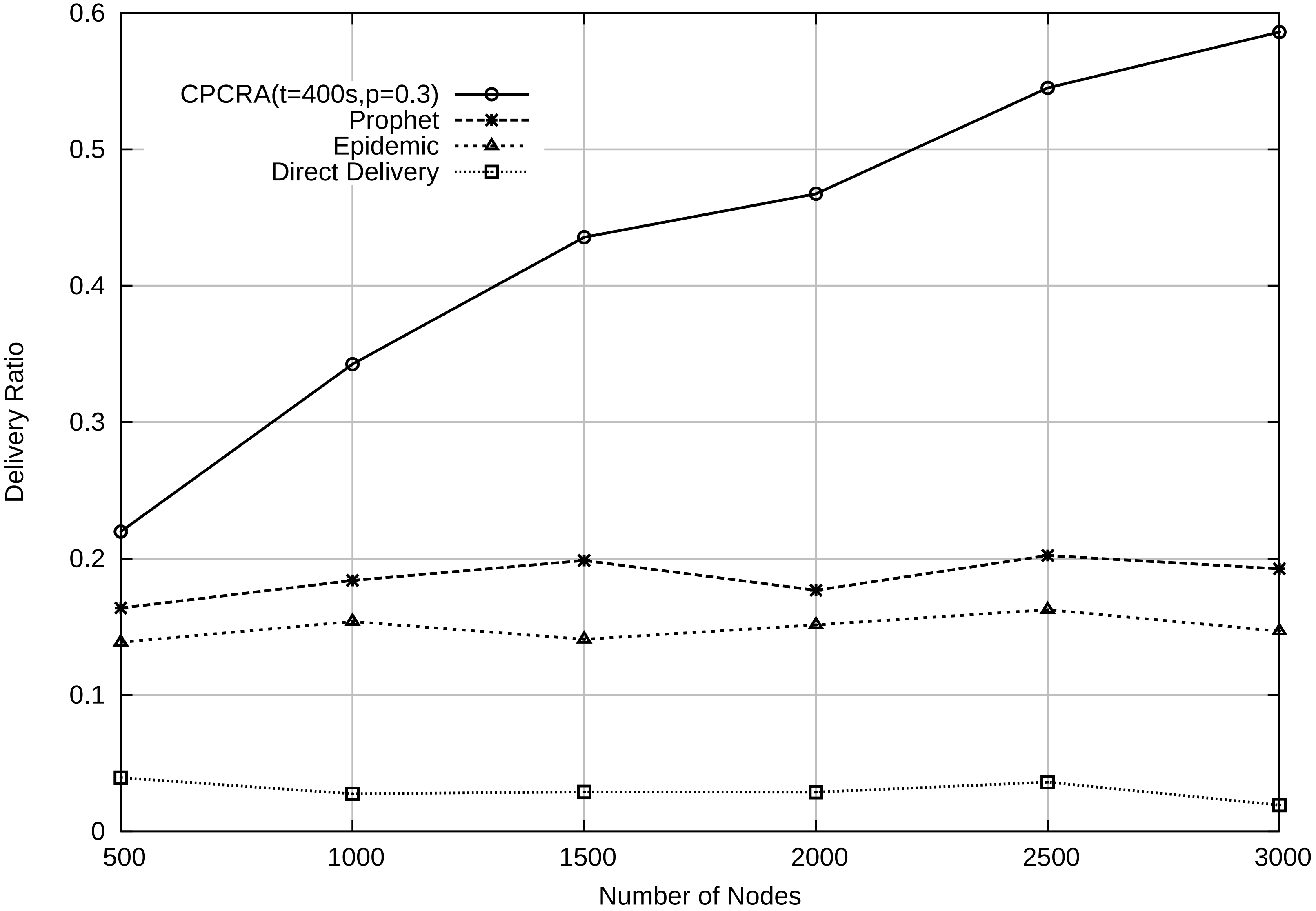
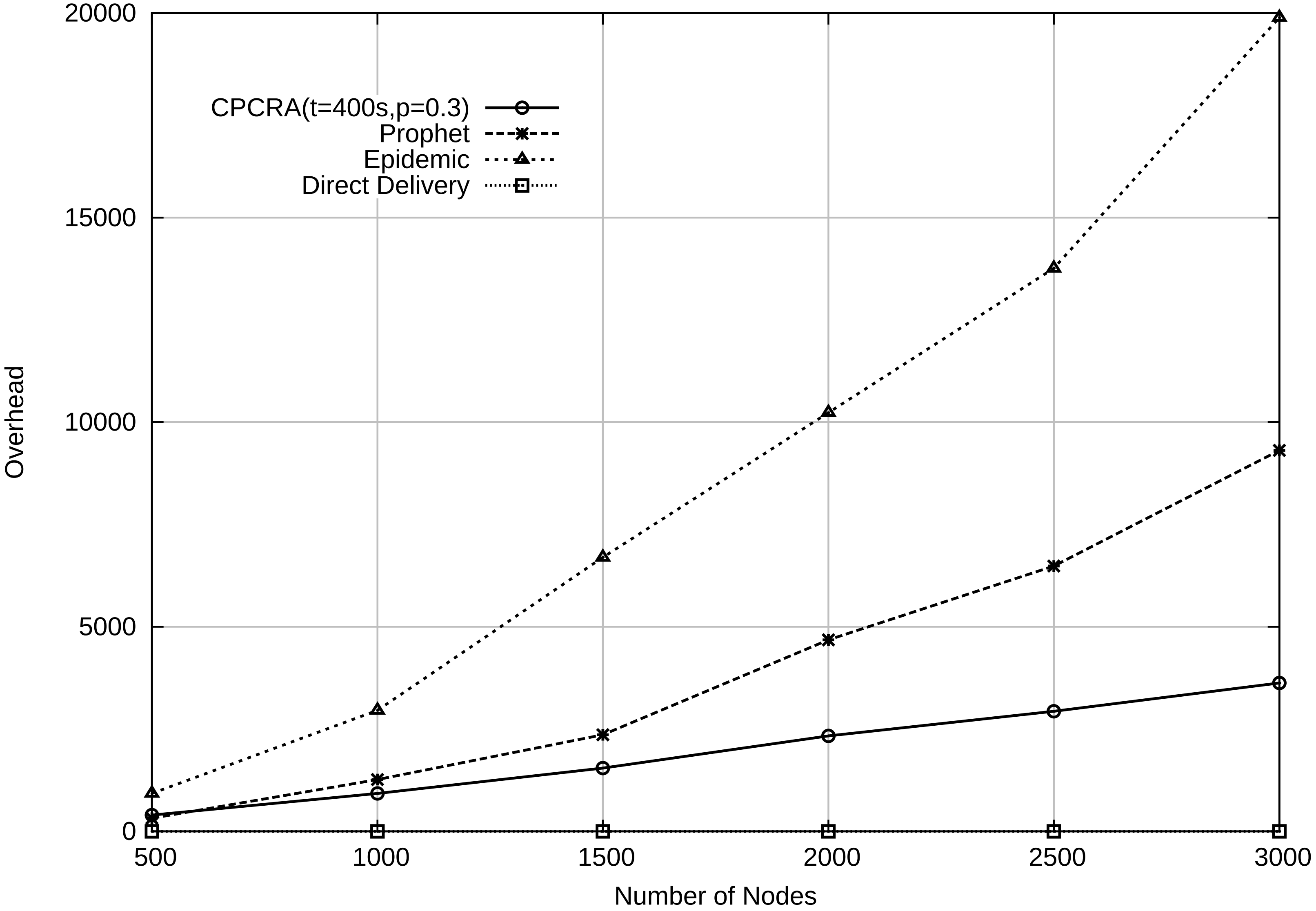


Fig. 14 Overhead of protocols with different network scales


CPCRA(t=400s,p=0.3)
Prophet
Epidemic
Direct Delivery
Overhead
Number of Nodes