Nodes Contact Probability Estimation Approach Based on Bayesian Network for DTN

Yuebin Bai*, Xu Shao,† Wentao Yang,* Weitao Wang*, Peng Feng*, Shuai Liu,* Xudong Zhang,* and Rui Wang*‡
*School of Computer Science and Engineering,
Beihang University, Beijing 100191, China
†Institute for Infocomm Research (I2R),
Agency for Science, Technology and Research, 609216, Singapore
‡Corresponding author, wangrui@buaa.edu.cn

***Abstract*—Delay tolerant network (DTN) known as suffering from frequent disruption, high latency and heterogeneous, resulting in low network availability. To improve DTN availability, routing protocols typically need to predict the probability of encountering the nodes. In this paper, we use the Bayesian Network (BN) to construct the knowledge base, which is an unique tool for creating a representation of the dependence relationships among DTN parameters. Then developed a Bayesian network-based approach to estimate the contact probability among nodes of DTN. We conducted an experiment to compare our approach against its counterparts in PROPHET routing protocol and power law distribution-based method. The experiment shows our approach is superior to other methods in both recall ratio and precision in all four datasets, including HAGGLE, NUS, REALITY and SASSY.**

***Index Terms*—Delay Tolerance Network, Bayesian Probability, Contact Probability Estimation**

I. Introduction

A Delay-Tolerant Network (DTN) is a network that has high end-to-end path latency, limited resources, frequent network partitions and so on characteristics[1],[2], which has many potential applications, such as Inter Planetary Internet (IPN), mobile vehicle network [3]. DTN research and development will strongly provide scientific theory and technical support for messages interaction of military war, aerospace communications, disaster recovery, emergency rescue and other fields and greatly promote the development trends of intellectualization, generalization and integration of network communication in the future.

The prediction of two nodes encounter probability is the core issue of DTN routing. For IPN and vehicular ad-hoc network (VANET) [5] with strong regularity and other network morphology, whose mobile law is comparative fixed and known, more precisely meet time of nodes encounter can be calculated by modeling the nodes motions, therefore, there is no need to predict encounter probability in such a network morphology. For portable switched network (PSN) [6], the mobile law of nodes can't be modeled, so that the nodes encounter probability is predicted mainly by dividing communities or directly calculating, according to which the routing makes its choice. This article will focus on researching the method of direct prediction of the probability of node encounters in the latter case.

978-1-5386-3416-5/18/$31.00 © 2018IEEE

In the encounter probability prediction method of PROPHET routing protocol [4], when one node has no contact with another, the probability of two nodes encountering will be gradually reduced over time; encounter probability prediction method based on the exponential distribution assumes that contact interval is exponentially distributed, and the predicted encounter probability is also gradually reduced over time.

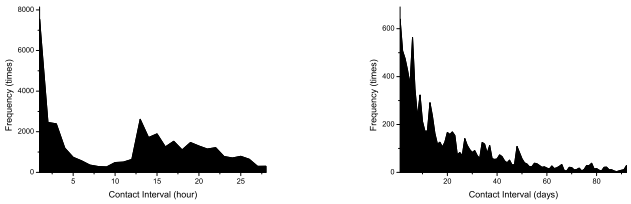
In order to better understand the problem, we analyzed two real DTN data sets, i.e. Haggle data set [7] and Reality data set [10], as is shown in $Fig.1$.

As can be seen from $Fig.1\ (a)$, at the beginning, with the growth of the contact interval, the frequency of contact interval (i.e. the probability of two nodes encountering again) is gradually reduced. In this case, the two before routing algorithm meet the actual situation. However, with the growth of time, when the contact interval is more than ten hours, the contact interval frequency will gradually increase with interval growth and peak in about 18- 24 hours. Combined with the scene of the data set collected, the time interval starting from the end of the first day meeting to the beginning of next day meeting is about 18-24 hours.

In $Fig.1\ (b)$, although the node contact time generally has a long-term regularity (i.e. contact probability reduced with the contact intervals growth), it is still can be seen that the distribution of the time interval is not a smooth curve, but there are many distinct peaks, and the peak arrange with a certain regularity: there will be a peak about every seven days or so. Combined with data set collected scene of view, there are strong week cycle regularity: courses, regular meetings, weekend social activities, campus activities, a lot of people will meet again in the interval of seven days or so.

Combined with the earlier analysis, it shows the encounter probability prediction method in PROPHET routing protocol and the encounter probability prediction method based on exponential distribution both are unable to give a accurate forecast.

In order to deal with this problem, we design a method called BACE to predict the probability of encountering the nodes. BACE can obtain this regularity by the statistics of



(a) Haggle Dataset (b) Reality Dataset

Figure 1: The Relationship between Contact Interval and Frequency.

historical data, and get the more accurate predictive value.

II. BAYESIAN NETWORK BASED APPROACH

In this subsection, we describe the Bayesian network based approach in detail. It is a challenging task to estimate the contact probability since there are many factors. We need to find the primary factor of contact pattern first. Firstly, we introduce the contact knowledge learning by creating a Bayesian network. Then we establish the approach based on this primary factor.

A. Contact Knowledge Learning by Bayesian Network

a) Candidate Factor Selection and Discretization

Node contact knowledge means the features that may have influence on the probability of two nodes met over a period of time, including contact frequency, average contact time, the average contact interval. Some of the factors are primary and not influenced by other factors, which is the factor we need to estimate contacts. In order to obtain the main factors affecting the nodes meet, we constructed a Bayesian network between these factors.

Frequency is the number of contact between two nodes in the history. With a higher frequency, two nodes may have a larger probability to contact each other in the future. Average contact time (ACT) is other measure of nodes familiarity.

Average contact interval (ACI) is a parameter that can be derived from device start-up time and ACT. Besides, node popularity (which means the number of nodes that a node had been encountered with) and node pair similarity (measured by number of common neighbor of a node pair) are also assumed to have great impact on node contact pattern.

b) Creating the Bayesian Network

A Bayesian Network(BN) is a graphical representation of relationships within a problem domain. More formally, a BN is a directed acyclic graph (DAG), where certain conditional independence assumptions hold. Bayesian network consists of a set of variables, a graphical structure connecting the variables, and a set of local conditional probability distributions. A Bayesian network is commonly represented as a graph, which is a set of vertices and edges. The vertices, or nodes, represent the variables and the edges, or arcs, represent the conditional dependencies in the model. The edges must be directed (the edges can be thought of as arrows) and there must be no cycles in the graph. It is often useful to think of the child nodes as being casually related to the parent nodes (the arrows are directed from parent nodes into a child node), although this does not necessarily have to be the case. In a Bayesian network, the joint probability distribution of all the nodes can be written as the product over all nodes of the conditional probability of each node given its parents. Structure learning is a procedure to construct the DAG. There are two methods are used mostly to design such DAG through a learning process known as structure learning [10].

The former is the constraint based method, in which a set of conditional independence statements is established based on some a priori knowledge or on some calculation from the data. Then this set of statements is used to design the DAG following the rules of d-separation. Another method,commonly used in the absence of a set of given conditional independence statements, is the score based method, which is able to infer a suboptimal DAG from a sufficiently large data set. This method consists of two parts: a function that scores each DAG based on how accurately it represents the probabilistic relations between the variables based on the realizations in the dataset and a search procedure that selects which DAGs to score within the set of all possible DAGs.

From the perspective of accuracy and efficiency, we adopt a very classic algorithm K2 which is the score-based method and celebrated for its high efficiency and accuracy.

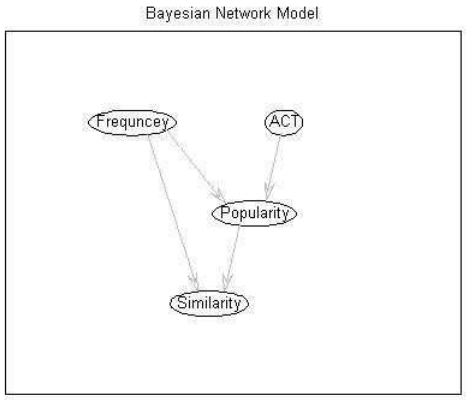


Figure 2: The Result of Structure Learning.

Through structure learning process, it is possible to build a model of the structure of a Bayesian network based on DTN factor variables, see $Fig.2$. As shown in the figure,the primary factors are the frequency of contact, the average contact time. Similarity and popularity are inferior factors affected by them. Bayesian network model illustrates the dependencies between the various network parameters, the probability relationship can be drawn between their specific parameter values. As we are going to estimating the contact probability, it's equivalent

to predict the average contact interval which can be derived from ACT, the primary factor. So our approach is going to be take consider of average contact interval as primary factor.

B. BACE:Bayesian Network Approach for Nodes Contact Estimation

Contact estimation refers to predicting the probability of contact in the next period of time of any two nodes, in the case of no contact, based on the historical data of the node information and the network environment.

To predict whether two nodes will be met within a certain period of time T_f, is to predict whether a node which had T_l time since last contact, will contact with the other nodes in the next T_f time. That is, whether the current contact interval I_c is shorter than T_f + T_l.

$$P(nodesencounterwithinT_f) = P(I_c < T_f + T_l) \quad (1)$$

As T_f + T_l are known, the whole problem is transformed to estimation of the length of the current interval Ic. As has already been waiting for the time of T_l, so $I_c \geq T_l$. So:

$$P(nodesencounterwithinT_f) = P(I_c < T_f + T_l | I_c > T_l) \quad (2)$$

By the conditional probability formula can be obtained:

$$P(I_c < T_f + T_l | I_c > T_l) = \frac{P(I_c < T_f + T_l \& I_c > T_l)}{P(I_c > T_l)} \quad (3)$$

Based on the formulas all above can be obtained:

$$P(nodesencounterwithinT_f) = \frac{f(T_l < I < T_f + T_l)}{f(T_l < I)} \quad (4)$$

In this equation, $f(x)$ stands for the frequency of x. In this equation, the frequency can be obtained through satiating historical data. In all, contact probability can be estimated with the available historical data.

III. EXPERIMENT AND ANALYSIS

A. Effectiveness Metrics

The experimental results can be divided into the four cases. As is shown in Table I, our prediction may give the right prediction about the contact (A1), and also we may have type I error (A2) or type II error (A3).

Table 1: PROBABLY RESULT FOR THE EXPERIMENT.

Actual Result	Estimating Results	
	Contact	No Contact
Contact	A1	A2
No Contact	A3	A4

The measurement of contact probability predict accuracy can refer to the scale used in information retrieval: recall ratio and check ratio, which are defined as follows:

$$RecallRatio = \frac{A1}{A1 + A2} \quad (5)$$

$$checkratio = \frac{A1}{A1 + A3} \quad (6)$$

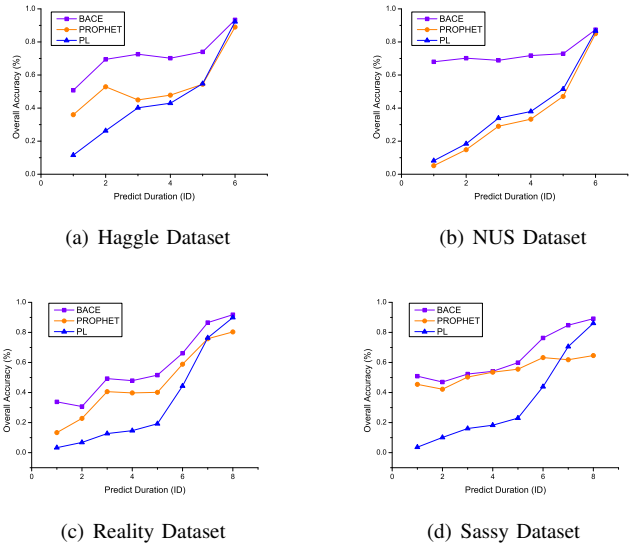
B. Universality of BACE

The method above have proved the availability of the Bayesian method and then compared the Bayesian method with other two methods in other four different types of datasets to verify its generality. The main parameters of the four data sets are shown in the following Table II. In this table, MCF and MCF are denoted as minimum contact frequency and minimum contact time, respectively.

Table 2:DATASET AND PARAMETERS FOR EXPERIMENT.

Parameters	Datasets			
	Haggle6 [7]	NUS[8]	Reality[9]	Sassy[10]
Node Number	41	3,000	91	25
Duration(s)	280,000	280,000	17,000,000	6,420,000
MCF	30	10	50	60
MCF(s)	50	3,599	20	5

After comprehensive consideration of recall ratio and check rate,comprehensive accuracy rate here is defined as geometric mean of the recall ratio and check rate, which can objectively reflect the prediction quality when the two rates above are much different. Taking the maximum comprehensive accuracy of each experiment as the representative of the prediction quality of this experiment, each group results are shown as follows.



(a) Haggle Dataset (b) NUS Dataset

(c) Reality Dataset (d) Sassy Dataset

Figure 3: Overall accuracy comparison among three methods.

As can be seen from the figure, BACE is better than the other two forecasting methods in various kinds of datasets and forecast period.

Results show that BACE performs the highest predictive accuracy for the three methods in all four datasets. The overall contact probability estimating accuracy of BACE have improved by 39.62% and 75.73% respectively compared to the PROPHET methods and Power-law distribution.

C. Routing Delivery Ratio

Experiments show that BACE is both effective and universal. In this section, we conduct a simulation to exam the performance of this method in actual scenario. The prove of improvement of effectiveness by statistic distribution is also presented in this section.

We implement our method in the widely-adopted DTN simulator ONE (opportunistic network environment simulator). The simulation parameters used in the two datasets are shown in Table 3. Meanwhile, each scenario is simulated and evaluated under three different buffer sizes: 30M, 50M and 100M.

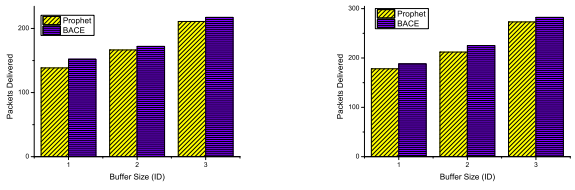
Table 3: SIMULATION PARAMETERS

	Haggle6	Reality
Update Interval (s)	1	3600
Simulation Duration(s)	350000	17000000
Nodes Number	98	97
Packets Number	330	561
Packets TTL(min)	500	30000
Buffer Size	[30,50,100]	[30,50,100]

In order to prevent errors caused by the random event simulation, several simulations with the same parameter are conducted. For each buffer size of each data set, ten rounds of simulations are repeated and the average number of packets successfully delivered is the standard measure of algorithm performance under the experimental conditions.

Simulation results of scenario employed Haggle dataset are shown in $Fig.4\ (a)$. It can be seen from the table that with the buffer increases, the number of successful packets delivered increases gradually. At the same time, under each buffer size, the BACE method always has a better performance on delivery rate.

Simulation results of Reality dataset are shown in $Fig.4\ (b)$. It can be seen from the table that with the buffer increases, the number of successful packets delivered increases gradually. At the same time, under each buffer size, the BACE method always have a better performance on delivery rate.



(a) Haggle Dataset (b) Reality Dataset

Figure 4: Delivery Ratio v.s. buffer size in ONE with Haggle Dataset and Reality Dataset.

The experimental results show that, on the different datasets, the BACE-based routing method have a better performance than PROPHET methods. At the same time, this improvements is more obvious with the smaller buffer size, which makes the improved routing methods have a better performance in the DTN with heave loads. Altogether, the the delivery ratio of BACE increase 5.18% than PROPHET method.

IV. Conclusion and Future Work

In this paper, we use the Bayesian Network (BN) to construct the knowledge base, which is an unique tool for creating a representation of the dependence relationships among DTN parameters. Then developed a Bayesian based approach to estimate the contact probability among DTN nodes. We conducted an experiment to compare our approach against its counterparts in PROPHET routing protocol and power law distribution based method. The experiment shows our approach is superior to other methods in both recall ratio and precision in all four datasets.For the future work, we are going to develop a routing protocol based on BACE.

Acknowledgment

This work is supported by the National Natural Science Foundation of China No.61572062, and No.61502019, the National High Technology Research and Development Program of China (863 Program) No. 2015AA016105, National Key Research and Development Program of China No.2016YFB1000503, Beihang University Innovation & Practice Fund for Graduate under Grant No.YCSJ-02-2017-03. The authors would like to thank great support.

References

[1] Kevin Fall. A delay-tolerant network architecture for challenged internets. In Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications . ACM, New York, NY, USA, Pages 27-34.

[2] Xiong Yongping, Sun Limin, Niu Jianwei, Liu Yan. Opptunity Networks. Journal of Software for China. 2009(01),Pages 124-137.

[3] Ian F. Akyildiz, Chao Chen, Jian Fang, Weilian Su. InterPlaNetary Internet: state-of-the-art and research challenges, Computer Networks, Volume 43, Issue 2, 7 October 2003, Pages 75-112.

[4] Anders Lindgren, Avri Doria, and Olov Schelen. Probabilistic routing in intermittently connected networks. SIGMOBILE Mobile Computing Communication. Rev.7,3.pp.239-254, July 2003

[5] Bret Hull, Vladimir Bychkovsky, Yang Zhang, Kevin Chen, Michel Goraczko, Allen Miu, Eugene Shih, Hari Balakrishnan, and Samuel Madden. CarTel: a distributed mobile sensor computing system. In Proceedings of the 4th international conference on Embedded networked sensor systems, pp.125-138 2006.

[6] Pan Hui, Augustin Chaintreau, James Scott, Richard Gass, Jon Crowcroft, and Christophe Diot. Pocket switched networks and human mobility in conference environments. In Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking,2005,244-251.

[7] Anirudh Natarajan, Mehul Motani, Vikram Srinivasan. Understanding Urban Interactions from Bluetooth Phone Contact Traces. PAM 2007, 8th Passive and Active Measurement conference. pp.115-124

[8] Nathan Eagle and Alex Pentland. Reality Mining: Sensing Complex Social Systems.Journal of Personal and Ubiquitous Computing. pp.255-268

[9] Greg Bigwood, Devan Rehunathan, Martin Bateman, Tristan Henderson, Saleem Bhatti. Exploiting Self-Reported Social Networks for Routing in Ubiquitous Computing Environments. Proceedings of IEEE International Conference on Wireless and Mobile Computing, Networking and Communication.484-489

[10] D Margaritis. Learning Bayesian Network Model Structure form Data, Ph.D. dissertation, School of Computer Science - Carnegie Mellon University, Pittsburgh, PA, May 2003.