Characterizing and Modeling Package Dynamics in Express Shipping Service Network

Xu Tan*, Yuanchao Shu*, Xie Lu†, Peng Cheng* and Jiming Chen*
*State Key Laboratory of Industrial Control Technology
Zhejiang University, Hangzhou, China
{xtan, ycshu}@zju.edu.cn, {pcheng, jmchen}@iipc.zju.edu.cn
†Department of Electrical Engineering and Computer Sciences
University of California, Berkeley, CA, USA
xielu@berkeley.edu

***Abstract*—Along with the increasing prosperity of market economy and the growth of online retail, express shipping service (e.g. FedEx, UPS) is playing an increasingly important role in our daily lives. A thorough understanding of the network structure and the package traffic dynamics of large-scale express shipping service network (ExpressNet) is essential for performance evaluation, network optimization, and user experience enhancement. Moreover, it would also be interesting and helpful to investigate how express shipping service reflects people's daily lives. In this paper, we propose systematic work to characterize and model the traffic dynamics in a nationwide ExpressNet. We collect 16 million delivery traces over 4 months in China, and examine its characteristics from a wide range of perspective, including network structure, temporal and spatial traffic dynamics, which provide important insights into express companies to better understand the network performance. On top of that, we develop an Extended Markov Model (EMM) to capture the dynamics of package delivery process and further predict the package delay, which is a major performance metric that both customers and express companies are concerned about. Data-based evaluation shows our model can achieve 91% prediction accuracy.**

I. Introduction

Due to the boom of online retail and increasing prosperity of market economy, express shipping service has become a major focus at the marketplace in recent years. For example, it is estimated that more than 9.1 billion packages have been delivered in 2013 through express shipping service network (ExpressNet) of China, which creates direct income around 254 billion RMB [1].

However, explosion in the number of packages brings new challenges to both package delivery companies and customers [2]. From the perspective of express companies, in order to provide better customer service, they need to design, evaluate and optimize the highly dynamic ExpressNet to help efficiently schedule the network resources [3]. On the other hand, customers need their packages to be delivered smoothly and timely. Although many delivery companies offer day-definite shipping service nowadays [4], they cannot provide accurate delivery time within a day beforehand and the delivery time of most express shipping services could even vary for a few days. With a better understanding and prediction of the package delivery dynamics, customers are able to estimate the shipping time more precisely and send out their packages at the right time and with the least costs.

Motivated by the above reasons, in this paper, we propose systematic work on characterizing and modeling package dynamics in ExpressNet. To this end, we firstly measure and characterize the temporal and spatial dynamics of express packages delivered within China in consecutive 4 months. After that, we model package delivery dynamics with an Extended Markov Model (EMM) and predict the package delay. In addition, we reveal some interesting correlations between ExpressNet and the socio-economic conditions and human behaviors, which can benefit the corresponding sociological studies.

Our study is distinguished from prior work in three aspects. First, we focus on analyzing and evaluating large-scale ExpressNet based on real data measurement. We design an effective data crawling algorithm and build our own data set using public package tracking data of Shunfeng Express, one of the largest express shipping service companies in China[1]. Such a data-driven design is more accurate and robust compared with previous analytical works in ExpressNet [5] [6] and the non-intrusive data collection method is flexible and easy to be implemented by ordinary users and third-party organizations for the purpose of research and evaluation. Second, unlike previous works which are mainly targeted on network structure design and optimization [7] [8] [9], based up on a huge data set, this paper provides thorough analysis of both temporal and spatial dynamics of package traffic with fine granularity. Specifically, our measurements and analyses provide hour-level time granularity and city-level spatial granularity, which can help express companies to better understand and further improve the performance of ExpressNet. Finally, to the best of our knowledge, we firstly propose a data-based model to predict package delivery time, which is beneficial for both express companies and

[1] http://www.sf-express.com/us/en/.

customers. Intellectual contributions of our work include:

1) From the public database of Shunfeng Express, we continuously collect package delivery traces for over 4 months and construct a data set including 16 million single traces which covers 284 cities in China. Based on the data set, we reconstruct the structure of ExpressNet, which is crucial to understand the unique topology characteristics and their influence on package dynamics.
2) We characterize several basic features of package dynamics in ExpressNet. More than 80% of packages are generated by only 20% of nodes. In addition, nearly 87% of packages are delivered within 2 hops while only 65.6% of packages follow the shortest path. We also discuss and reveal the underlying reasons of the observations.
3) The time-series of the whole traffic volume show obvious diurnal characteristics and have a strong relation with people's daily activities. We further demonstrate how the package delivery delay is largely dependent on the randomness from these characteristics in both temporal and spatial domains. The insights obtained from the analysis can help express companies to further understand the network performance.
4) We utilize the measurement analysis results to develop the EMM to capture the package delivery dynamics. Through large-scale data set based evaluation, EMM shows high accuracy in package delay prediction.

The remainder of this paper is organized as follows. In Section II, we describe the data set used in our work and reconstruct the network structure. In Section III, we reveal the fundamental features of package traffic dynamics on the spatial and temporal scale. Then the EMM is developed to capture the package delivery dynamics in Section IV. We discuss past work in Section V and along with a conclusion of this paper in Section VI.

II. Preliminary

In this section, we first provide a quick sketch and discussion of the collected data set. On top of that, we reveal some structure features of ExpressNet, which are essential for later analysis and modeling of the package dynamics.

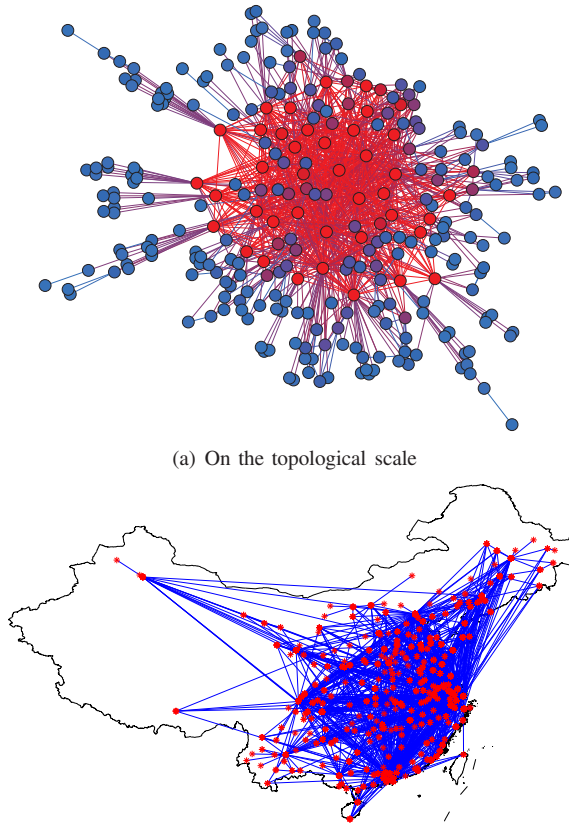
A. Data Set Description

We collect the package tracking data from the open website of Shunfeng Express that provides nationwide package delivery services in China. With hour-level temporal granularity and city-level spatial granularity, such tracking data can reveal the delivery trace and dynamics of the package, including when and where the package is collected, sorted, conveyed and delivered. Notice that the data trace does not contain any private or sensitive information such as the name or the contact number of the customer, the exact location where the package is picked up, and etc. Therefore, collecting and analyzing the data will not bring in any privacy issues. In our work, we develop a specific application that is able to crawl such package delivery traces automatically. The collected data set contains 16 million package traces from November 2012 to February 2013, including 284 major cities throughout China.

There are also limitations of the data set. The recorded time in each trace may not accurate due to the delay of package scanning process. For example, packages arrived early at a hub are probably scanned at a later time. In order to estimate the possible errors, we sent out hundreds of packages containing sensor nodes to obtain the ground truth time and location information. The measurements show that although the scanning time varies within a batch of packages, the variance is small enough, usually less than an hour, to be neglected in our analysis and modeling in this paper.

B. ExpressNet Structure

In this part, we reveal the structure of ExpressNet by analyzing the collected data traces in the view of network science [10]. The results contribute to the understanding of the topology characteristics and their influence on package delivery process.



(a) On the topological scale

(b) On the geographical scale

Figure 1. ExpressNet structure

The ExpressNet structure abstracted from the traces is shown in Figure 1. Nodes in Figure 1 represent cities in collected traces, and edges represent delivery links between two nodes. Figure 1(a) is the general view of ExpressNet structure on the topological scale, where the degree of nodes in the center is larger than that in the outer layer. Figure 1(b) shows the actual geographical location of ExpressNet in China.

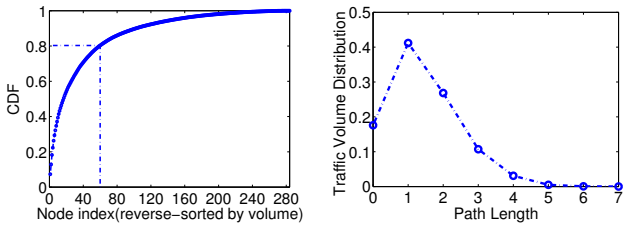
The ExpressNet we collected contains 284 nodes, 1253 edges and 10196 paths. Specifically, a path is a distinctive package delivery route between the source node and the corresponding destination node (node pair) which may have many traces spanning over the observed time period. The average node degree is 8.824 and the average path length is 2.704.

III. Characterizing Package Traffic Dynamics

In this section, we study the fundamental characteristics of the package traffic in ExpressNet. We first measure the traffic volume from both spatial and temporal domains and then analyze the package delay. The insights gained from the interesting traffic characteristics are not only of significant importance for service management, schedule planning and network optimization, but also can benefit the customers' daily lives.

A. Traffic Volume Distribution

Traffic volume is defined as the number of packages that counted from the collected package traces. Figure 2(a) shows the cumulative distribution function (CDF) of traffic volume on different nodes. It can be seen that nearly 20% of nodes own 80% of traffic volume, indicating that traffic volume in ExpressNet is dominated by a small part of nodes. Figure 2(b) shows the traffic volume distribution on different path length. The single-hop path bears more than 40% of the whole traffic volume. Besides, almost 20% of traffic volume belongs to the path with the length of 0, which reveals the importance of intra-city package traffic.



(a) Reverse-sorted distribution of traffic volume on nodes

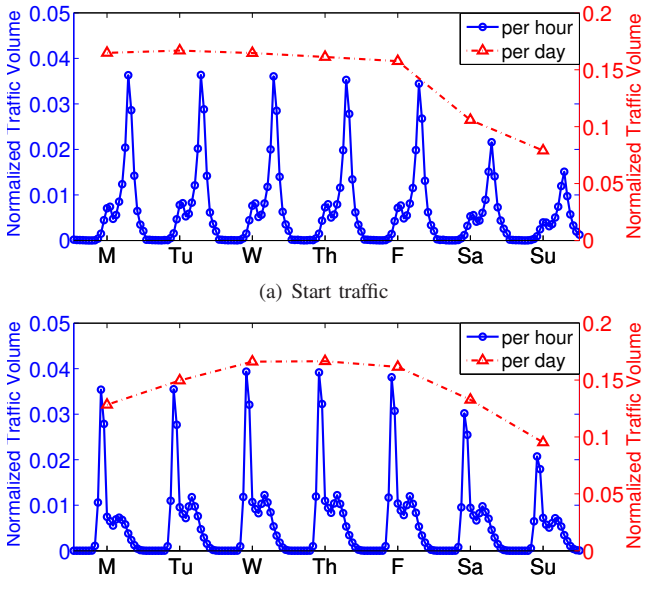
(b) Traffic volume distribution on different path length

Figure 2. Traffic volume distribution on nodes and paths in ExpressNet

In many other networks such as communication networks, the shortest paths are usually chosen for information exchange. However, it's not the case in ExpressNet. The proportion of the traffic flow on all the shortest paths (the shortest path is defined as the path with the fewest hops between node pair) is 65.6%. It indicates that more than one path may exist between a pair of nodes and a great part of traffic volume does not follow the shortest path rule. This can be explained by the unique features of ExpressNet that package delivery is carefully scheduled and the delivery time is usually strictly limited [3]. In the case when the trucks scheduled over the shortest path have left already, the network may figure out another path in order to meet the deadline [9]. Therefore, different paths between nodes will clearly complicate the package delivery dynamics and add difficulties to the delay analysis as we will discuss in later sections.

B. Traffic Temporal Dynamics

It is interesting to measure the traffic dynamics on the temporal scale that owns hour-level granularity. In this part, traffic generated from the source node (start traffic) and the traffic flowing to the destination node (end traffic) is analyzed and compared. We first observe a clear periodicity of the start and end traffic on one-week scale at per-hour granularity. Then we shrink to one-day scale and extend to one-month scale, respectively, to make comparison between the start and end traffic.



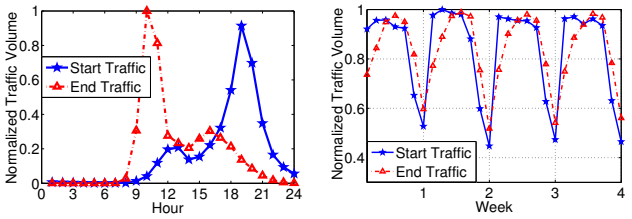
(a) Start traffic

(b) End traffic

Figure 3. Diurnal characteristics of start and end traffic volume

Figure 3 shows the weekly distribution of start and end traffic volume. Firstly, we focus on the traffic time-series at per hour granularity (the blue circle dot lines in Figure 3(a) and 3(b)). The package volume in both start and end traffic concentrates in the day time and reaches the peak at a relatively fixed time every day, i.e., they show strong

diurnal characteristics. More clear details about start and end traffic are compared on one-day scale in Figure 4(a). For the start traffic, the peak is around 19:00 and this indicates that packages collected in the daytime are mostly sent out in the evening especially between 18:00 and 20:00. There is also a small peak around 12:00, which corresponds to some priority packages collected in the morning and mailed at noon. For the end traffic, the peak arrives at 10:00 and a slight shake exists between 15:00 and 17:00. It indicates that the majority part of packages arrive at their destination in the morning and are expected to be delivered by noon. Another circumstance is that packages arrive between 15:00 and 17:00. Such a schedule is also reasonable because the couriers of express company will be usually off duty after 19:00.



(a) One-day scale (b) One-month scale

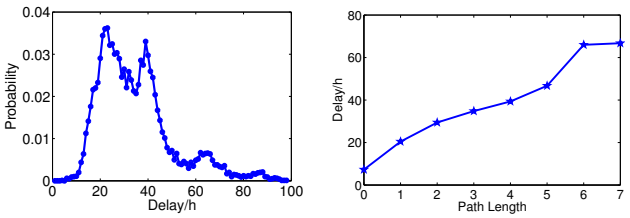
Figure 4. Comparison between start and end traffic

Furthermore, we discuss the traffic volume time-series at per day granularity (the red triangle lines in Figure 3(a) and 3(b)). In Figure 3(a), traffic counts from Monday to Friday are relative stable and are higher than that during the weekend, which of the pattern is consistent with working schedule in China. Most packages are generated to satisfy the demand of the business activities on weekdays. Comparatively, less express service is needed during weekends. It's interesting to notice that the red triangle lines in the two figures do not follow the same trend. Thus we compare the time-series between start traffic and end traffic on the one-month scale in Figure 4(b). The results reveals that most of the end traffic lags 1-2 days compared with the start traffic. This time shift corresponds to the average delivery time, which we will discuss in the Section IV.

C. Package Delay Analysis

In this section, we characterize an important property of traffic dynamics: the package delay. Package delay represents the delivery time between package pickup and package receipt. The delay distribution and the relationship between delay and distance are analyzed. The results obtained from the delay analysis further motivate us to model the package delivery dynamics in Section IV.

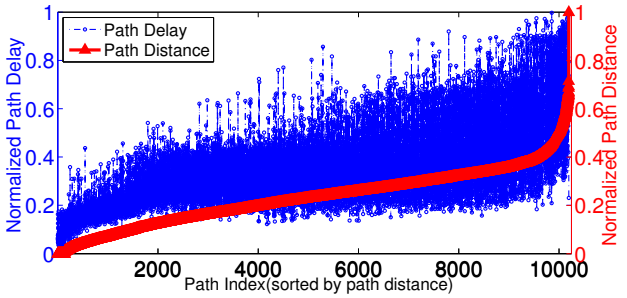
We find that package delay varies a lot on paths with different distances. Even packages on the same path may have a diverse delay for the different start time and transportation methods. The distribution of the package mean delay on all paths are shown in Figure 5(a). The delay varies from 1 to 98 hours and the majority of delay is less than 50 hours. Moreover, two peaks appear near 23 and 39 hours, which are in accordance with the "next-day delivery" and "two-day delivery" services provided by Shunfeng Express.



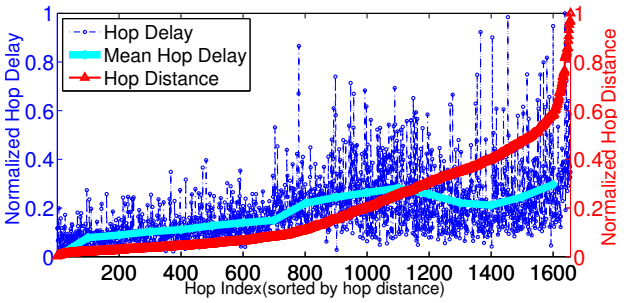
(a) The distribution of the mean delay on all paths (b) The mean delay on different path lengths

Figure 5. The distribution of the mean delay

In order to discover the relationship between delay and path length, we plot the mean delay on different path length in Figure 5(b). It clearly shows that the mean delay increases with the path length. This is similar to the intuition that more hops usually add up distance, which results in a larger delay. However, it's difficult to predict the package delay with this relationship because it is just a statistical phenomenon and does not contain accurate distance information.



(a) The relation between path delay and the corresponding distance



(b) The relation between hop delay and the corresponding distance

Figure 6. The relation between delay and distance

To be more reasonable, we take the geographic distance into consideration and plot the relationship between mean delay and path distance in Figure 6(a). The blue dotted line

represents the normalized mean delay on different paths (the path is sorted by distance) and the red delta line shows the normalized path distance. The delay increases with the path distance in the overall trend. But this trend is not obvious enough and even some short delays appear when the path distances are relatively long. We believe this is caused by different transportation methods in delivery process.

To verify this hypothesis, we analyze the relationship between mean delays and hop distances in Figure 6(b). The delay first increases with the hop distance. But when the hop index (the hop is sorted by hop distance) reaches 1100 (the corresponding hop distance is 859 kilometers), the delay begins to decrease. It implies when hop distance is longer than 859 kilometers, airplanes are used thus short delay appears. It can be further supported by the major peak delay near 23 hours in Figure 5(a). The threshold hop delay corresponding to the threshold distance is 16.8 hours, adding the delay inside the nodes together up to 23 hours. It's obvious that such a threshold distance has been carefully designed. It should not be too long for the trucks to miss the time limit while not too short as more airplanes incur high costs.

D. Summary

In this section, we have presented thorough analyses on package traffic dynamics in ExpressNet. Traffic volume is uneven distributed across different nodes and different lengths of paths. 20% of nodes own 80% of traffic volume and single-hop path bears more than 40% of traffic volume. Different paths may exist between a pair of nodes and a large part of traffic volume does not follow the shortest path rule, which complicates the package dynamics. These analysis results provide useful insights for service management and network optimization, such as node importance estimation or flow control in ExpressNet.

The package traffic shows strong diurnal pattern and weekly variation, which reflects people's daily activities. Traffic usually starts at nightfall as transportation during night is fast and cost-efficient. It also guarantees more packages collected in the daytime be delivered at at an early time of a day, which is consistent with the fact that traffic often ends in the morning. In addition, traffic volume of weekday is larger than that during the weekend indicating the weekly working schedule.

There are two peaks in the mean delay on all paths, which are in accordance with the two types of services provided by Shunfeng Express. A positive correlation exists between package delay and path length as well as path distance. We hypothesize that 859 kilometers may be a threshold in the statistical sense for the company to use airplane and further verify it through concrete analysis. The dynamics in the package delivery is relatively complicated and the package delay is hard to precisely infer through data analysis. This motivates us to investigate how to model the package delivery dynamics mathematically using our collected data traces, thus predicting the delay more accurately.

IV. Modeling Package Delivery Dynamics

Delivery time/delay is a key performance metric in express shipping service. To estimate the delivery delay, we need to sketch the network dynamics. From Section III-A and III-C, we find estimation of package delay is a non-trivial problem due to the following two reasons. The first one is that multiple paths may exist between a pair of nodes and packages do not always take the shortest path. The other one is that different kinds of transportation methods such as trucks and airplanes may be used on the path. To solve this problem, in this section we propose a data-based Extended Markov Model to depict the package delivery dynamics and further predict the package delay. The basic model is described in Section IV-A and then the modeling performance is evaluated in Section IV-B. In Section IV-C, we discuss how to utilize the model to estimate delay between nodes on different paths.

A. Extended Markov Model

For ExpressNet $G = (V, E)$, the state of package delivered in G can be represented as $s(l)$, where l is the location of the package and can either be the node v or the edge e in the delivery path. Taking the path $R = v_1v_2v_3$ as an example, all the possible states in this path are $s(v_1)$, $s(v_1v_2)$, $s(v_2)$, $s(v_2v_3)$ and $s(v_3)$. When describing a delivery process, the probability of the package in a certain state at a given time needs to be concerned. Therefore, we define $X^k = [x_1\ x_2\ ...x_i\ ...\ x_n]$ as the state probability set of the package delivered in the path with n states at time slot k. x_i is the probability of the package in the $i-th$ state, $0 \leq x_i \leq 1$, $\Sigma_{i=1}^{n} x_i = 1$. For the path $R = v_1v_2v_3$ mentioned above, if the state probability set $X^k = [0.2\ 0.3\ 0.3\ 0.2\ 0]$, it means that the probabilities of the package of the 5 states are 0.2, 0.3, 0.3, 0.2 and 0, respectively, at time slot k. Therefore, the package delivery process can be defined as a random process $X =< X^0, X^1, X^2, ...X^k, ... >$. X can be modeled as a discrete time Markov chain as follows:

$$X^k = X^{k-1}P \tag{1}$$

where P is the transition matrix. Distinguished from conventional transition matrix, P is changing over time in the delivery process. We term this kind of model as Extended Markov Model (EMM) in this paper. Mathematically, the EMM can be described as follows:

$$X^k = X^{k-1}P^k = X^0 \prod_{i=1}^{k} P^i \tag{2}$$

where $X^0 = [1\ 0\ 0\ ...\ 0]$ is the initial state probability set and P^i is the transition matrix at time slot i. The process will terminate when X^k reaches $[0\ 0\ ...\ 0\ 1]$.

The EMM can not only depict the package dynamics during the delivery process, but also predict the package

delay. However, there are two factors required to be derived. One is the length of each time slot, i.e. the period T of the delivery process, and the transition matrix P^k. The period T is set to 1 hour. Such a setting should be accurate enough for the time granularity while reducing the computational complexity of P^k. Then we describe how to compute P^k.

For a path with n states, P^k can be wrote as follows:

$$P^k = \begin{bmatrix} {p_{11}}^k & {p_{12}}^k & 0 & 0 & ... & 0 \\ 0 & {p_{22}}^k & {p_{23}}^k & 0 & ... & 0 \\ ... & ... & ... & ... & ... & 0 \\ ... & ... & {p_{ii}}^k & {p_{ij}}^k & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & 0 & {p_{nn}}^k \end{bmatrix} \quad (3)$$

where p_{ij}^k denotes the transition probability from state x_i to x_j at time slot k, $\Sigma_{j=1}^{N} p_{ij}^k = 1$. For a given time slot k, the package can either transfer to the next state or stay at the original state. Therefore, $p_{ii}^k = 1 - p_{ij}^k$. Then we just need to compute p_{ij}^k in P^k.

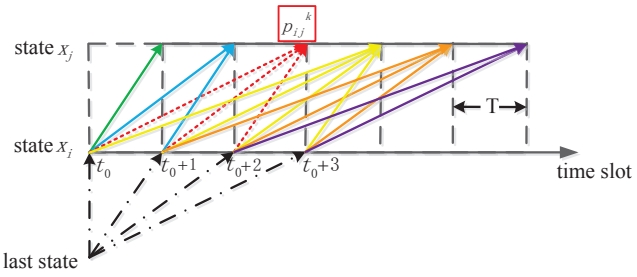


Figure 7. The transfer process between two states

For a given path $R = v_1 v_2 ... v_i v_j ... v_n$, we just consider two states $s(v_i)$ and $s(v_i v_j)$. The state probability is denoted as x_i, x_j, respectively. In the following parts, x_i, x_j can also represent the states $s(v_i)$ and $s(v_i v_j)$. The transition process from state x_i to x_j is shown in Figure 7. t_0 is the earliest time slot that package arrives at x_i (It also represents the package collection time if x_i is the first state in the path). The probability of package transferring to state x_j from x_i depends on the time slot when package arrives at x_i. The red dotted line in the figure is taken as an example to describe the deduction of p_{ij}^k. Mathematically, p_{ij}^k can be computed as

$$p_{ij}^k = \sum_{m=0}^{k-1} \widetilde{Pr}(t_0 + m) Pr_{t_0+m}(k - m - 1) \quad (4)$$

where $\widetilde{Pr}(t_0 + m)$ is the probability that package arrives at x_i at time slot $t_0 + m$, and $Pr_{t_0+m}(k - m - 1)$ is the probability that package transfers to x_j at time slot $k-m-1$ under the condition that package arrives at x_i at time slot $t_0 + m$. Both $\widetilde{Pr}(t_0 + m)$ and $Pr_{t_0+m}(k - m - 1)$ can be calculated from the corresponding traces in our data set.

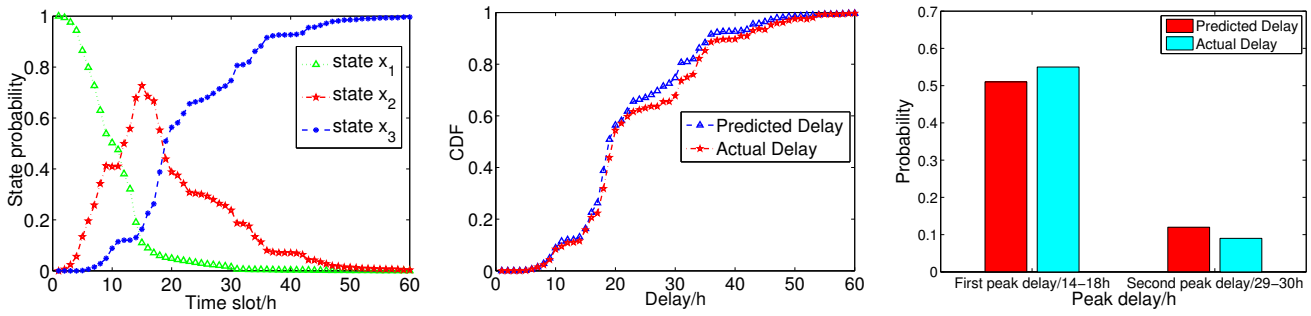
B. Evaluation

In order to understand the performance of the EMM, the package dynamics and the delay prediction obtained by the model are tested and validated in this section. According to the data generated time, the data set is divided into two parts. The first part consists of the data from November to December in 2012, which is used as the training data to calculate model parameters. The second part contains the data from January to February in 2013 and is used for modeling performance evaluation. In the rest of this section, the model is firstly tested on an easy single-hop path to show how to utilize the model. Then the model is implemented in a more complicated multi-hop scenario to further validate the performance. At last, various multi-hop path cases are randomly selected to verify the accuracy of delay prediction.

First we analyze the package dynamics and delay prediction on a single-hop path from Beijing to Shanghai. Figure 8(a) shows the changes of the state probability set in this path. It can be observed that state x_1 decreases from 1 to 0 and x_3 increases from 0 to 1. Meanwhile, state x_2 increases in the first 15 hours and decreases to 0 later. It indicates that package is more likely to transfer from node Beijing to hop Beijing-Shanghai in the first 15 hours, and then transfer from hop Beijing-Shanghai to node Shanghai. The distribution of x_3 predict the package delay at the same time. For instance, x_3 is 74.53% at time slot 30, which represents the probability of delay less than 30 hours is 74.53%.

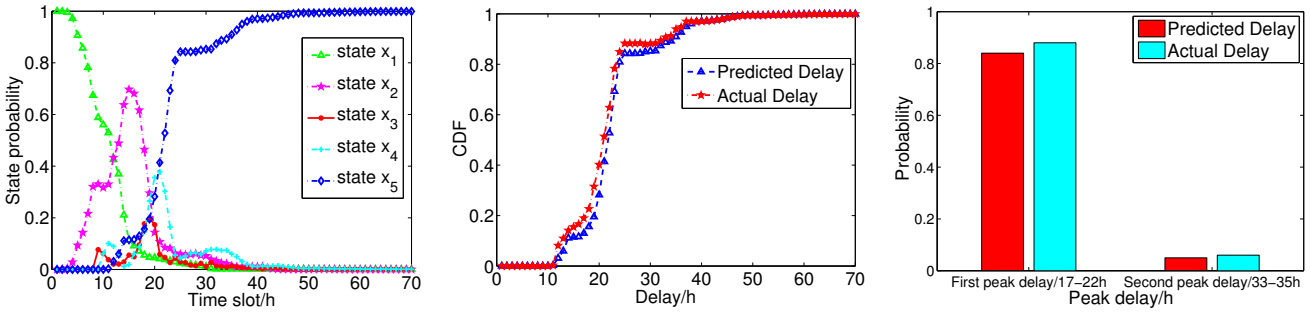
Figure 8(b) compares the CDF between the delay predicted by the model and the actual delay obtained from the evaluation data set. It can be observed that two CDF curves match well. Great decreases occur in both curves when the delay is close to 16 hours, which is a potential delivery time of the package. The comparison of the delay with peak probability is illustrated in Figure 8(c). It can be seen that the predicted delay with maximum probability is between 14 and 18 hours, with probability as high as 51%. While the actual probability based on the evaluation data set is 55%, with a probability difference of only 4%. The probability of predicted delay and actual delay between 29 and 30 hours is 12%, 9%, respectively, with a deviation of around 3%. The analysis above demonstrates that our model works well for the single-hop case.

Figure 9 shows the evaluation results of the model on the multi-hop path from Beijing to Guangzhou, via Shenzhen. It can be seen that x_5 in Figure 9(a) is 83.65% at time slot 25 indicating the probability of package delivery time less than 25 hours is 83.65%. Figure 9(b) further shows the well consistency of the CDF between the predicted delay and the actual delay. Peak probability of delay is shown in Figure 9(c). The probability of predicted delay between 17 and 22 hours is 84% while the ground truth value is 88%. These evaluation results validate our delay prediction model for the multi-hop case.



(a) State set (b) CDF comparison (c) Peak value comparison

Figure 8. The evaluation of the single-hop case



(a) State set (b) CDF comparison (c) Peak value comparison

Figure 9. The evaluation of the multi-hop case

We further evaluate the EMM with different multi-hop paths in our data set. It can be seen from Figure 2(b) that less than 40% of the total traffic volume is multi-hop transportation, which demonstrates that traffic volume is unevenly distributed among paths. Actually a great part of multi-hop paths have little traffic volume that is not sufficient for modeling. Therefore, we select multi-hop paths with plenty of package traces for further evaluation, which include 200 paths with totally 3.4 million traces (21% of the total traces). As analyzed above, the delay on most paths owns a peak value with the maximum probability, so we define the model accuracy δ as

$$\delta = 1 - \frac{\|Pr_{pre} - Pr_{act}\|}{Pr_{act}} \tag{5}$$

where Pr_{pre} and Pr_{act} are probabilities of the predicted peak delay and the actual peak delay, respectively. After testing all 200 paths, we find the average value of δ achieves 91%, which verifies the effectiveness of our proposed EMM in package delivery delay prediction.

C. Discussion

After the validation of the model accuracy, we further describe the utilization of the model to predict the delay between node pair with different paths. For a given node pair S and D, suppose there exist n different paths $R_i, i \in [1, n]$ between them. From the evaluation section above, we can see the peak delay of each path can be obtained from the EMM, denoted as $P^k(R_i)$. It's reasonable to use the peak delay to depict the delay on the path because most paths in our data set own a peak delay with a dominant probability. A simple way to calculate the peak delay between this node pair is

$$P^k(SD) = \sum_{i=1}^{n} \beta_i P^k(R_i) \tag{6}$$

where β_i is the weight of each path and can be simply determined by the proportion of traffic flow on this path. The coefficient β_i is a adjusted parameter which could be used to characterize path uncertainties and various transportation methods between nodes. We leave the problem of how to choose β_i to our future work.

Note that based on the collected data set spanning over 4 months, our EMM provides fairly high accuracy for different node pairs and path length. However, the structure of ExprressNet may change over time. In that case, we need to update the model parameters periodically using the latest collected data therefore maintain a high prediction accuracy.

V. Related Work

To the best of our knowledge, this paper presents the first study on characterizing and modeling the package dynamics in a ExpressNet based on real package delivery

information. Prior related works include the network design and optimization, the geographical analysis of the network and the impact of express service on economy. We briefly describe some primary related works below.

In [7], the authors present a linear programming formulation for the single allocation p-hub problem using data from a postal delivery network. In [8], *Seung-Ju Jeong et al.* investigate how to plan the transport routes, frequency of service and transportation volume in rail freight system in a 10-country European network. *Kim et al.* develop a model for package delivery problems with time windows in [9]. In contrast to these works on network design and optimization, our work focuses on the characterization and performance evaluation on the ExpressNet put in operation through extensive data analysis. The results from our work can in turn evaluate and verify the validity of the design.

In addition, there are also some works analyzing the ExpressNet from the perspective of social policy and economics. *C.-C. Lin et al.* compare the economic effects of hub-and-spoke network with center-to-center network in [5]. *Y.-C. Song et al.* analyze the relationship between the logistics and economic growth in China based on the data from National Bureau of Statistics in [6]. Different from these works, we not only characterize the traffic dynamics but also model the package delivery process in a relatively rigorous way, to benefit both service providers and customers.

There are several related works using the similar methods to study the traffic in other networks. In [11] [12] [13] [14], the authors characterize the traffic dynamics in cellular networks. Relevant to our work, they measure the temporal traffic dynamics based on network science. The authors in [15] present a model to combine multiple random processes in network traffic, and in [16], the authors model the dynamic trust of online service providers by a Hidden Markov Model. Similar to our work, both of them model the intricate traffic dynamics. However, our data-based EMM is more robust and flexible, and can self-updated with newly collected data. Some other related works include characterizing the video traffic of YouTube [17] and wireless network traffic during the Super Bowl [18].

VI. Conclusion

In this paper, we propose complete systematic study on characterizing and modeling package dynamics in express shipping service network (ExpressNet). Based on 16 million collected express shipping traces, we first infer the network structure, and then examine the spatial and temporal characteristics of the network. Furthermore, we devise an Extended Markov Model to depict the package delivery dynamics and predict the package delay. The performance of the model is extensively evaluated with high delay prediction accuracy. We believe the results and insights gained from our work can promote express shipping service benefits both service providers and customers.

Acknowledgment

This paper was partially supported by NSFC under Grants 61222305, 61190110, National Program for Special Support of Top-Notch Young Professionals, and Fundamental Research Funds for the Central Universities under Grant 2014XZZX001-03, 2014XZZX003-25.

References

[1] S. P. B. of China, "The postal industry development statistics bulletin 2013," http://www.spb.gov.cn/dtxx_15079/201401/t20140115_274540.html.

[2] ResearchInChina, "China express delivery industry report," http://www.researchinchina.com/Htmls/Report/2013/6772.html.

[3] C. Barnhart, N. Krishnan, D. Kim, and K. Ware, "Network design for express shipment delivery," *Computational Optimization and Applications*, vol. 21, no. 3, pp. 239–262, 2002.

[4] S. Express, "Product overview," http://sf-express.com/cn/en/product_service/product_intro/express_delivery.html.

[5] C.-C. Lin, Y.-J. Lin, and D.-Y. Lin, "The economic effects of center-to-center directs on hub-and-spoke networks for air express common carriers," *Journal of Air Transport Management*, vol. 9, no. 4, pp. 255–265, 2003.

[6] Y.-C. Song and W. Lv, "Research on data mining of logistics and economic development," in *International Conference on Computational Intelligence and Software Engineering*, 2009.

[7] A. T. Ernst and M. Krishnamoorthy, "Efficient algorithms for the uncapacitated single allocation p-hub median problem," *Location science*, vol. 4, no. 3, pp. 139–154, 1996.

[8] S.-J. Jeong, C.-G. Lee, and J. H. Bookbinder, "The european freight railway system as a hub-and-spoke network," *Transportation Research Part A: Policy and Practice*, vol. 41, no. 6, pp. 523–536, 2007.

[9] D. Kim, C. Barnhart, K. Ware, and G. Reinhardt, "Multimodal express package delivery: A service network design application," *Transportation Science*, vol. 33, no. 4, pp. 391–407, 1999.

[10] S. Chen, W. Huang, C. Cattani, and G. Altieri, "Traffic dynamics on complex networks: a survey," *Mathematical Problems in Engineering*, vol. 2012, 2011.

[11] L. Qian, B. Wu, R. Zhang, W. Zhang, and M. Luo, "Characterization of 3g data-plane traffic and application towards centralized control and management for software defined networking," in *IEEE Big Data*, 2013.

[12] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *ACM SIGMETRICS*, 2011.

[13] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "A first look at cellular machine-to-machine traffic: large scale measurement and characterization," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 40. ACM, 2012, pp. 65–76.

[14] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *IEEE INFOCOM*, 2011.

[15] M. Laner, P. Svoboda, and M. Rupp, "Modeling randomness in network traffic," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, 2012, pp. 393–394.

[16] X. Zheng, Y. Wang, and M. A. Orgun, "Modeling the dynamic trust of online service providers using hmm," in *IEEE International Conference on Web Services (ICWS)*, 2013.

[17] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: a view from the edge," in *ACM IMC*, 2007.

[18] J. Erman and K. Ramakrishnan, "Understanding the super-sized traffic of the super bowl," in *ACM IMC*, 2013.