On Understanding Diffusion Dynamics of Patrons at a Theme Park

(Extended Abstract)

Jiali Du[†] Akshat Kumar Pradeep Varakantham
[†] Living Analytics Research Center
School of Information Systems
Singapore Management University, Singapore, 178902
{jiali.du.2012, akshat.kumar, pradeepv}@smu.edu.sg

ABSTRACT

In this work, we focus on the novel application of learning the diffusion dynamics of visitors among attractions at a large theme park using only *aggregate information* about waiting times at attractions. Main contributions include formulating optimisation models to compute diffusion dynamics. We also developed algorithm capable of dealing with noise in the data to populate parameters in the optimization model. We validated our approach using cross validation on a real theme park data set. Our approach provides an accuracy of about 80% for popular attractions, providing solid empirical support for our diffusion models.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed AI—*Multiagent Systems*

General Terms

Algorithms, Theory

Keywords

Collective graphical models, Diffusion models

1. INTRODUCTION

Diffusion dynamics refers to how entities spread in an underlying network. Understanding mobility pattern in a theme park is interesting for several reasons. First, it helps predict and control the contagion spread in a multiagent system. Secondly, understanding this mobility pattern is attractive as better strategies can be adopted to ease the overall congestion at different attractions. However, due to prohibitive cost of instrumentation, obtaining individual movement is difficult. Therefore, diffusion dynamics must be learned from *aggregate* or *collective* data about the underlying flow.

Reasoning with noisy aggregate data is an emerging research area. Collective graphical models (CGMs) are recently developed as a general framework for reasoning with aggregate data in the context of a probabilistic graphical model [2]. Our work is a special case of CGMs. The key differentiator in our work is that we exploit the structured dynamics of visitor movement to develop simpler and tractable optimization-based formulations of the diffusion dynamics problem. Furthermore, we apply our techniques to the real-world theme park problem, which has not been done previously.

Recently, Kumar *et al.* [1] formulated the problem of learning turn-probabilities at road intersections in a traffic network based on aggregate information about vehicle inflow and outflow. Our work addresses an enriched version of this problem as flow conservation is more complex in a theme park owing to the presence of queues at each attraction. A significant contribution is to formulate and validate diffusion models with real world data, which is not provided in [1].

We consider a real theme park in Singapore, where the problem of severe congestion has been observed consistently over the last few years. Since congestion is primarily associated with major attractions, we learn diffusion models for the 9 major attractions. For achieving this goal, we have access to a 5-month long data set of wait times throughout the day for all the attractions.

2. VISITOR DIFFUSION MODEL

We use a layered time indexed graph to represent the flow dynamics. Each layer and node represents one time slice and attraction, respectively. We denote the service rate at an attraction i in a single time interval using s_i. Service rate typically depends on the nature of attraction. Bold letters $\boldsymbol{n}$,$\boldsymbol{x}$ and $\boldsymbol{p}$ are used to represent vectors composed of all individual elements. $n_{d,t,i}$ denotes the number of visitors waiting to be serviced at node i, time t in cascade d. Similarly, $x_{d,t,i,j}$ corresponds to the number of people moving from node i to j; $p_{t,i,j}$ represents the probability that a visitor would move from node i to j.

In the *Observation Model*, the goal is to learn the parameters $\boldsymbol{p}$ from observations $\boldsymbol{x}$. In reality, it is impractical to get the vector $\boldsymbol{x}$. In this work, we focus on *Partial Observation Model.* In this model, we focus on learning the underlying diffusion model using only aggregate observation $\boldsymbol{n}$.

Three diffusion models are followed showing how their parameters can be learned based on partial observations.

Multinomial Distribution Based Diffusion: The likelihood of the complete data $\boldsymbol{x}, \boldsymbol{n}$ is given as:

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

Variables: $\boldsymbol{p}, \boldsymbol{x}$

$$\text{Maximize: } \sum_{d}\sum_{i}\sum_{t}\Big(\log\big((\sum_{j} x_{d,t,i,j})!\big) - \sum_{j}\log(x_{d,t,i,j}!) + \sum_{j} x_{d,t,i,j}\log(p_{t,i,j})\Big)$$

Subject to:

$$n_{d,t+1,i} = n_{d,t,i} + \sum_{k} x_{d,t,k,i} - \sum_{j} x_{d,t,i,j} \quad \forall d,t,i \qquad (3)$$

$$\sum_{j} x_{d,t,i,j} \leq min(s_i, n_{d,t,i}) \quad \forall d,t,i \qquad (4)$$

$$\sum_{j} p_{t,i,j} = 1 \quad \forall t,i \qquad (5)$$

$$x_{d,t,i,j} \in \mathbb{N}_0 \quad \forall d,t,i,j \qquad (6)$$

$$0 \leq p_{t,i,j} \leq 1 \quad \forall t,i,j \qquad (7)$$

Table 1: GetDiffusionDynamics($\boldsymbol{n}$, $\boldsymbol{s}$)

$$\mathcal{L}(\boldsymbol{p};\boldsymbol{x},\boldsymbol{n}) = \prod_{d\in D}\prod_{i\in A}\prod_{t\in T}\frac{(\sum_j x_{d,t,i,j})!}{\prod\limits_{j\in A} x_{d,t,i,j}!}\prod_{j\in A} p_{t,i,j}^{x_{d,t,i,j}} \qquad (1)$$

where D denotes the observed cascades, A denotes the set of all attractions and T is the set of time slices. Total outflow visitors of node i at time t is $\sum_j x_{d,t,i,j}$. This corresponds to the total number of trials in the multinomial distribution. Each subsequent attraction represents the number of possible outcomes in the multinomial distribution.

To compute $\boldsymbol{p}$, we maximize the likelihood and use following approximation to make it computationally simpler:

$$\max_{\boldsymbol{p},\boldsymbol{x}} \log P(\boldsymbol{n},\boldsymbol{x};\boldsymbol{p}) \qquad (2)$$

The above optimization problem can be formulated as a nonlinear program shown in Table 1. Objective function is the log of Eq. (1). The first and the second constraint jointly represent the flow conservation at each node. The rest enforce some basic properties of the diffusion model. We use Lingo to solve the optimization problem in Table 1.

Dirichlet-Multinomial Based Diffusion: The Dirichlet-Multinomial distribution provides a prior $\boldsymbol{\alpha}$ for the generation of the diffusion model. Instead of constraints (5) and (7) on $\boldsymbol{p}$, we used a lower bound of 2 and upper bound of 4 for $\alpha_{t,i,j}$ values.

Poisson Distribution Based Diffusion: Similarly, the key parameter of interest in addition to $\mathbf{x}$ with Poisson distribution is $\boldsymbol{\lambda}$. The optimization formulation in Table 1 is appropriately modified to reflect this change.

3. EVALUATION

We use a 5-month long data set of wait times at a theme park to evaluate our approaches. To account for lack of data on visitors entering and exiting the theme park as well as taking breaks, we introduce a new attraction called the 'leisure' node. This attraction is numbered as 'A10' with infinite service rate and capacity. As we maximize the likelihood, our formulation in Table 1 ensures the diffusion model has the fewest possible population as transitioning to and from the 'leisure' attraction. Moreover, the 'leisure' attraction also has the positive effect of accounting for errors in reporting of wait times without violating the flow conservation constraints.

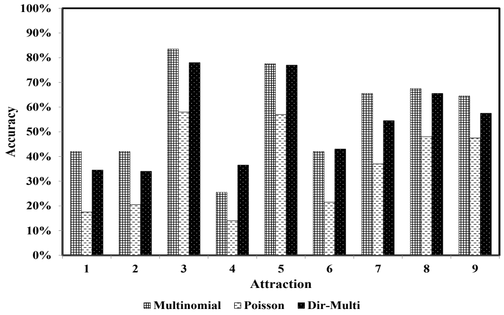

Figure 1: Accuracy with respect to the attractions

3.1 Results by Cross Validation

In order to understand the accuracy of transitions predicted using the formulation in Table 1 and others, we apply 5-fold cross validation with following steps:

- We first solve the formulation in Table 1 on the training data set to obtain the underlying parameters $\boldsymbol{p}$ ($\boldsymbol{\alpha}$ for Dirichlet-multinomial and $\boldsymbol{\lambda}$ for Poisson) of the distributions corresponding to each attraction.
- We then use the parameters obtained by solving the optimization problem to assess the accuracy.
- By considering fixed confidence intervals, we count the total accuracy of prediction at each time step.

Accuracy results are provided in Figure 1 with respect to individual attractions for the 30% confidence intervals. From the figure 1, multinomial distribution consistently provides higher accuracy than the other two distributions. A key insight is that the Poisson distribution performed significantly worse. Our result indicates that Poisson distribution may not be ideal to represent a network of queues, where the status of one queue depends on the status of other queues. One practically important observation is that attractions that have high wait times (3, 5, 8 and 9) also have a higher accuracy of prediction (70%-87%).

4. ACKNOWLEDGMENTS

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

5. REFERENCES

[1] A. Kumar, D. Sheldon, and B. Srivastava. Collective diffusion over networks: Models and inference. In *International Conference on Uncertainty in Artificial Intelligence*, pages 351–360, 2013.

[2] D. Sheldon, T. Sun, A. Kumar, and T. Dietterich. Approximate inference in collective graphical models. In *International Conference on Machine Learning*, pages 1004–1012, May 2013.