Research Statement

Amin Vahedian Khezerlou

My primary research interests fall into the fields of data mining and machine learning. Specifically, my research contributions are in the area of Urban Big Data Analytics (UBDA). Urban big data is any type of data that records observations of an urban phenomenon, such as human mobility, air quality, Points of Interest (POI), etc in a scale that covers a significant portion of a metropolis. I believe urban big data provides an unprecedented opportunity to advance our capabilities, both as citizens and governments, in achieving important goals such as urban sustainability, economic growth and more. I define UBDA as the computational analysis of urban big data with the purpose of realizing such goals.

My work in UBDA has mostly focused on the analysis of rare patterns of gathering and dispersals in human mobility. The emphasis on studying *rare* patterns is an important distinction of my work. Regular patterns are less challenging to formulate, detect and/or predict than rare patterns, because of the inherent imbalance of the data in favor of regularity. I develop formulations and methods that detect and/or predict rare gatherings and dispersals using real-time mobility data such as real-time traffic flows, GPS traces and taxi pickup-drop records. My contributions facilitate better handling of challenges that arise as a result of rare gathering and dispersal patterns, such as traffic jams and public safety concerns as wells as creating business opportunities for a range of services such as transportation and advertising.

My other work in this field is through collaborations with teams that are working on a variety of challenging problems such as prediction of traffic congestion propagation, optimizing taxi revenue based on optimal passenger seeking strategies and understanding business location patterns. These collaborations have broadened my expertise and general view of scientific research practices through professional interactions with researchers with a diverse set of backgrounds.

As I will present later in this statement, my research is a combination of real-world problem solving and technical contributions. I am interested in novel computational formulations of real-world problems that have not been addressed before. I am particularly excited by the computational challenges that arise in such formulations and the trade-off between solution quality and computational cost. Maintaining a balance between applied and theoretical aspects of research is a core principle of my work.

1 Thesis Research

As mentioned earlier, my dissertation formulates concepts and develops methods that facilitate the mining of urban big mobility data. Specifically, the aim of the formulations and developed methods is to identify and predict certain events that occur as a result of urban mobility. I study unexpected gathering and dispersal events.

1.1 Gathering Events

A Gathering event is the process of an unusually large number of moving objects (e.g. taxi) arriving at the same area within a short period of time. It is important for city management to identify emerging gathering events which might cause public safety or sustainability concerns.

Prior work to detect gathering events uses undirected patterns which lack the ability to specify the dynamic flow of the traffic and the destination of the gathering. I propose the concept of a Gathering Graph (G-Graph) to model the footprint of a gathering event by including both the gathering destination and the paths to the destination. Early detection of gathering events is challenging due to numerous candidate gathering footprints in a spatial field and the non-trivial task to balance pattern quality and computational efficiency. I propose the SmartEdge Algorithm for early detection of gathering events. An algorithm that uses innovative pruning and indexing strategies to efficiently detect gathering events and build their G-Graphs. Figure 1 shows the effect of the three design decisions on improving the running time compared to a brute-force baseline, by increasing the size of the studied spatial region. The fast running time of the algorithm allows it to run at every time-step and detect top gathering events based on real-time counts. Figure 2 shows an example of top 5 gathering events at a time-step in a spatial field defined for the study area in Shenzhen, China. This portion of my dissertation led to a publication in ACM SIGSPATIAL'16 conference [8] and a journal publication in ACM Transactions on Intelligent Systems and Technology (TIST) [2].

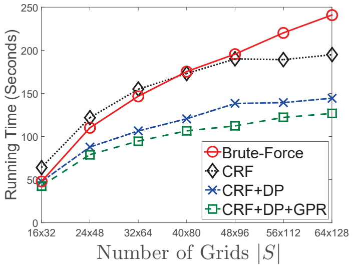


Figure 1: Running time performance of the SmartEdge Algorithm.

Forecasting gathering events is a predictive approach as opposed to descriptive approaches of detection. For the first time, I use destination prediction to forecast gathering events to take a predictive approach to analysis of gathering events. Destination prediction is challenging due to complex dependencies among the segments of each trajectory. This challenge has been addressed in the literature by modeling the trajectory as transitions between locations, which are treated as Markovian states [5]. Formulating the trajectory as a Markov process enforces the inherent assumption of independence from the past. In the context of an urban trip, this assumption is severely limiting, because future locations of a traveler strongly depend on its past locations.

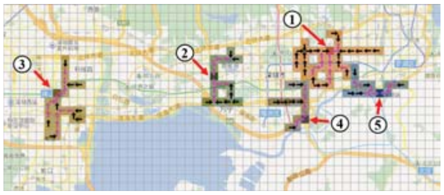
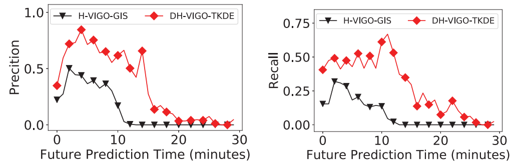


Figure 2: An example of detected G-Graphs.

I relax this limiting assumption and address the resulting computational challenge by proposing a state-of-the-art destination prediction model called Via Location Grouping (VIGO) that efficiently produces destination probabilities for incomplete trajectories. The memory and time efficiency of VIGO is obtained by taking advantage of spatial partitioning and smart indexing strategies.

Even by using an accurate destination predictor, it is challenging to forecast unexpected gatherings. Because, learning historical pattens of trajectories cannot reliably forecast rare gathering events, as they violate regular patterns by behaving abnormally. To address this challenge, I propose a Dynamic Hybrid framework, called DH-VIGO, that takes advantage of two VIGO models. This framework is capable if identifying and learning emerging patterns of abnormality in the trajectories, as well as historical patterns. DH-VIGO dynamically decides which pattern (historical vs. recent abnormality) should be used to predict the destination of each incomplete trip. DH-VIGO makes it possible to forecast rare and unexpected gathering events *ahead of the time*, by pinpointing to the scarce and not-immediately-obvious evidence that exists at current time. Figure 3 shows the prediction performance of DH-VIGO in forecasting gathering events in future time-steps compared to a baseline I developed earlier in my dissertation. This portion of my dissertation resulted in a publication in ACM SIGSPATIAL'17 conference [3] and a journal article submitted to IEEE Transactions in Data and Knowledge Engineering (TKDE), which was recently revised.

1.2 Dispersal Events



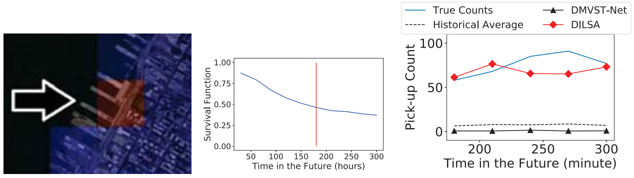
(a) Precision. (b) Recall.

Figure 3: Prediction performance of DH-VIGO.

A dispersal event is the process of an unusually large number of moving objects leaving the same area within a short period of time. Early prediction of dispersal events is important in mitigating congestion and safety risks and making better dispatching decisions for taxi and ride-sharing fleets. Literature of dispersal event prediction solves this problem as a taxi demand prediction problem. It is shown that taxi demand has a highly predictable pattern [7]. However, dispersal events are by definition violations of this predictable pattern. Thus, existing methods fail to give accurate predictions of demand in case of dispersal events. There are three main questions in this study: (1) Will there be a dispersal event in the future? If so, (2) when is the starting time? And (3) what is the demand during the dispersal event?

To answer these questions I propose a two-step framework by formulating the Dispersal event prediction problem as a Survival Analysis problem. I call this framework DILSA. The proposed framework is capable of predicting the occurrence and start time of dispersal events in addition to the abnormal demand in case of such an event. In my formulation, the occurrence of dispersal event is treated as the *death* event in conventional survival analysis. This formulation uses deep artificial neural networks to estimate the survival function for a target period in the future. Using DILSA, future dispersal events are predicted with recall of 0.9 and precision of 0.6. The average error of predicting the event's starting time is 18 minutes, using 30-minute time-steps.

Figure 4 shows an example of a real-world dispersal event predicted by DILSA on March 19, 2016 at Pier 92/94 in Manhattan as a result of a home design exhibition. Figure 4 (b) shows the survival function during the prediction target period. The vertical line shows the predicted start time of the dispersal anomaly. Figure 4 (c) shows the predicted pickup counts during the dispersal event. The figure shows that DILSA predicts the unexpected high demand, while the baseline DMVST-Net [6] stays close to the historical average. Based on this work I have submitted a paper to the AAAI 2019 conference, which is currently under review.



(a) Event location. (b) The Survival curve. (c) Predicted pick-up counts vs. true counts.

Figure 4: An example of a predicted dispersal event (best viewed in color).

2 Other Research

I participated in a research project that recommended optimal passenger seeking routes to taxi drivers using a data-driven approach. Given GPS traces of taxis, which include the location history of each taxi in both passenger-carrying and passenger-seeking modes, we developed a model based on Markov Decision Process that recommends the best next move to the driver, while they are seeking passengers. The important distinction of this work with its literature is that it optimizes the recommendations based on not only the pick-up probability, but with overall profit that takes the likely destinations of the potential passengers into account. This project led to a journal publication in IEEE Transactions on Big Data [9].

I also collaborated in a project to predict traffic congestion propagations. In this study, we presented a computational formulation of congestion propagation and proposed a prediction model based on historical conditional probabilities of propagation. A preliminary draft of this work is to be presented in IWCTS'18 [4].

In another collaboration, I mentored a young researcher through the University of Iowa SSTP program to analyze the business location patterns in major metropolitan areas of the United States. Through this study we understand the significant co-location patterns of businesses inside or across the boundaries of industries. For instance, we find that major gas stations are significantly declustered in New York City, while local gas stations tend to cluster. We present many such analysis that aim to provide an objective picture of the business location choice strategies. A preliminary draft of this work is to be presented in 2^{nd} INFORMS Workshop on Data Science 2018 [1].

3 Future Research

In short-term, I plan to expand on the dispersal events analysis. I plan to advance the analysis to the following stages of the dispersal events. Once a dispersal event occurs, the traffic flowing away from the dispersal location also carries the attribute of abnormality and poses its own challenges to urban mobility. I plan to formulate the aftermath of the dispersal to demonstrate the most likely impacted paths out of the dispersal location. That is, the paths that experience more-than-expected traffic flow. I plan to investigate the best predictors of those paths, while anticipating the historical trajectories to be bad predictors. Balancing the computational efficiency and the quality of the patterns will be another challenging aspect of this project.

In another short-term plan, I address urban gatherings with a different point of view. In this plan, I would like to answer the question: given a location and time, what are the consequences of having a large gathering at that location and time? By *consequences*, I specifically mean, what road segments will experience congestion. There is an important distinction between this question and the early-detection and forecasting problems in the sense that, here I ask a *what if* question as opposed to *when* and *where* questions. To answer this question, I plan to develop a model that will predict the resulting congestions given a gathering and its volume at a location and time. Developing such a model is challenging, because it must consider where people will come from and what routes they will take. Given rare nature of the gatherings, these patterns are challenging to obtain from historical data.

In medium term, I plan to develop a comprehensive framework that gives an answer to a bigger *what if* question: Given a set of events and their magnitudes, what are the best locations for them within an urban area, that will minimize the *cost* of all events. The cost, is a comprehensive measure that will be developed to include, gathering impact, dispersal impact, cost of modification to existing road segments or constructing new ones, cost of modification to existing venues or constructing new ones. The input for this framework would be a spatial field that includes the road network and information of available real estate and the cost of acquiring it, plus information on modification and new construction costs and historical trajectories within the spatial field. Such a comprehensive formulation is challenging not only because of its broad scope, but also because of the large size of the solution space.

Beyond Urban Big Data Analytics. While I keep myself closely involved in UBDA, I maintain a consistent line of thinking on the generalization potential of all the developed methods and formulations to the broader field of data mining and machine learning. We live in interesting times. Today, data is being gathered with a different mindset than decades ago. It is not about what needs to be recorded, it is about what *can* be recorded. Ongoing improvement of this *ability-to-record* has quickly pushed us to a point where manual examination of data is infeasible. Data mining and machine learning are results of a particular human invention, the computer, and the desire to achieve the idealistic goal of *understanding* the data. While *data* gives information on the world we live in, we very often seek its exploitation over its understanding. To this end, we delegate the task of *understanding* to the machines. While this approach is most often justified, I disagree with a full delegation. I believe, while we enjoy the success of our black box predictors, we must ultimately move towards understanding the phenomenon recorded by the data, by building machines that are capable of teaching us their understanding of it.

My tendency towards *what if* questions is a result of perusing this long-term goal. For instance, this consideration occupied my mind while I was developing VIGO, the destination prediction model. In essence, VIGO is a data structure that implements Bayes' optimal classifier by keeping track of certain counts that can be used for instantaneous probability calculations. As a result, it can be built to produce joint probability distributions of any set of features conditioned on another set of features, which provides remarkable potential for interpretability of an already accurate model. VIGO is not limited by any assumptions of dependency, or lack there of, among features. The only limitation of VIGO is its computational cost, specifically the memory cost, which was handled in my project by taking advantage of certain attributes of urban trips.

I plan to drive my career using these broad principles. I will consider my career successful if I am able to move towards my larger goals through my short and medium-term contributions.

References

[1] Chiu, J., Vahedian, A., and Zhou, X. Understanding business location choice pattern: A co-location analysis on urban poi data. In *2nd INFORMS Workshop on Data Science* (2018), INFORMS.

[2] Khezerlou, A. V., Zhou, X., Li, L., Shafiq, Z., Liu, A. X., and Zhang, F. A traffic flow approach to early detection of gathering events: Comprehensive results. *ACM Transactions on Intelligent Systems and Technology (TIST) 8*, 6 (2017), 74.

[3] Vahedian, A., Zhou, X., Tong, L., Li, Y., and Luo, J. Forecasting gathering events through continuous destination prediction on big trajectory data. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2017), ACM, p. 34.

[4] Xiong, H., Vahedian, A., and Zhou, X. Predicting traffic congestion propagation patterns: A propagation graph approach. In *11th ACM SIGSPATIAL International Workshop on Computational Transportation Science (IWCTS'18)* (2018), ACM.

[5] Xue, A. Y., Qi, J., Xie, X., Zhang, R., Huang, J., and Li, Y. Solving the data sparsity problem in destination prediction. *The VLDB Journal—The International Journal on Very Large Data Bases 24*, 2 (2015), 219–243.

[6] Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., Gong, P., Ye, J., and Li, Z. Deep multi-view spatial-temporal network for taxi demand prediction. In *2018 AAAI Conference on Artificial Intelligence (AAAI'18)* (2018).

[7] Zhao, K., Khryashchev, D., Freire, J., Silva, C., and Vo, H. Predicting taxi demand at high spatial resolution: Approaching the limit of predictability. In *Big Data (Big Data), 2016 IEEE International Conference on* (2016), IEEE, pp. 833–842.

[8] Zhou, X., Khezerlou, A. V., Liu, A., Shafiq, Z., and Zhang, F. A traffic flow approach to early detection of gathering events. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2016), ACM, p. 4.

[9] Zhou, X., Rong, H., Yang, C., Khezerlou, A. V., Zheng, H., Shafiq, Z., and Liu, A. X. Optimizing taxi driver profit efficiency: A spatial network-based markov decision process approach. *IEEE Transactions on Big Data* (2018).