Mining Big and Fast Data: Algorithms for Large-Scale Data Processing

Muhammad Anis Uddin Nasir

May 31, 2017

1 Current Research

In the last decade, most of the research and industry has been focused on processing big and fast data as massive amount of data has been produced on numerous hand-held devices due to high bandwidth connectivity. Also, most of the data is produced at a very rapid rate, making old information less relevant, for instance, Facebook users are continuously creating new connections and removing old ones, thus changing the network structure. Twitter users produce posts at a high rate, which makes old posts less relevant and changes, for instance, the retweet and mention networks. Developing analytical tools to process this amount of information at a fast speed is challenging, yet extremely essential, for developing new services in the areas such as web analytics, e-health and marketing.

My main research interests revolve around designing scalable and data-driven algorithms for real-time linked-data analysis (i.e., streams and graphs). Such algorithms enable mining latent patterns from massive and dynamic data (i.e., logs, web, and networks). The main goal is to mitigate the challenges for various systems and applications. One of my major research contributions is the scalability study for distributed stream processing engines. In particular, we studied the the load-balancing problem for distributed stream processing engines, which is caused by the skewness in the workload and heterogeneity in the cluster. In doing so, we developed three different algorithms to reduce the load imbalance across a distributed system. Our results are well accepted by both the research and open-source communities[6, 7, 5, 4]. Moreover, few of our algorithms are integrated into Apache Storm, which is an open source stream processing framework. Further, we designed a data driven algorithm

for distributed hash tables to improve the network performance [3]. We showed that such distributed hash tables drastically improve the performance of several network related applications.

Apart from the research related to systems, I am immensely interested in studying various problems related to data mining, data management and machine learning. My goal is to design algorithms for computationally challenging (NP-hard) problems that enable answering difficult questions. Alongside, we studied the top-k densest subgraph problem in a fully-dynamic settings and proposed an extremely efficient algorithm for the problem that scales to graphs with billions of edges [2]. Morevoer, we presented efficient algorithms for streaming graph partitioning problem [8, 1]. Currently, I am studying the influence maximization problem in the fully-dynamic setting, which is one of the most-studied problems in the last decade and plays an important role in the domain of viral marketing.

Lastly, I have been actively collaborating with researchers from several renowned organizations, such as Yahoo Labs, Telefonica Research, IBM Research, Qatar Computing Research Institute, Swedish institute of Computer Science and Aalto University. I feel that collaborating with experienced researchers from various fields have broaden my research experience and enabled me to look at algorithmic problems from different perspectives.

2 Areas of Strength

My areas of strength are parallel to my research that focus on designing algorithm for various challenging problems. My expertise is divided into several domains: distributed systems, data mining and machine learning. I would like to focus more on these domains and help to bridge them in order to produce novel and efficient algorithms for important problems.

3 Areas for Development

I am aiming to build a strong theoretical knowledge that is extremely important for modern research in the field of computer science. I have already taken several offline and online courses related to advanced algorithms, combination optimization, graph theory, approximation algorithms, and mathematics for computer science. Moreover, I have attended several summer schools in the domain of algorithms for large scale

distributed systems. However, I still feel that there is a need to put more effort to specialize in this domain. I believe getting strong foundation in the domain of theoretical computer science will help me in achieving my research goals in a proficient manner. Further, I want to apply for several academic grants including both individual and joined grants. My goal is to collaborate with my colleagues within and outside consortium to put together joint grant proposals targeting European funding agencies.

4 Awards and Funding

I have received several awards and funding throughout my academic career. One of my major fundings has been a 3-year Marie Curie Scholarship for PhD studies at KTH Royal Institute of Technology. Also, I have received Erasmus Mundus scholarship for European Master in Distributed Computing as well as various academic scholarships during my undergraduate studies at National University of Science and Technology (NUST). Lastly, I have been accepted for several internships during my academic career.

5 Planning For Your Career Development Goals

I have two immediate career goals. First, I aim to finish my PhD studies in the first quarter of the year 2018. During the final stage of my PhD, I target to get more publications in the top venues in the field data mining and management. Second, I aim to pursue postdoctoral research in a strong and ambitious group that enables me to work on challenging problems. My motivation to join a strong group can be summarized in this famous quote by Henry Ford: "If everyone is moving forward together, then success takes care of itself". Finally, I would like to expand my network of collaborators to polish my existing skills and develop new skills that might help me to advance in my research career.

References

[1] M. A. U. Nasir. Gossip-based partitioning and replication middleware for online social networks, 2013.

[2] M. A. U. Nasir, A. Gionis, G. D. F. Morales, and S. Girdzijauskas. Top-k densest subgraphs in sliding-window graph streams. *arXiv preprint arXiv:1610.05897*, 2016.

[3] M. A. U. Nasir, S. Girdzijauskas, and N. Kourtellis. Socially-aware distributed hash tables for decentralized online social networks. In *Peer-to-Peer Computing (P2P), 2015 IEEE International Conference on*, pages 1–10. IEEE, 2015.

[4] M. A. U. Nasir, H. Horii, M. Serafini, N. Kourtellis, R. Raymond, S. Girdzijauskas, and T. Osogami. Load balancing for skewed streams on heterogeneous cluster. *arXiv preprint arXiv:1705.09073*, 2017.

[5] M. A. U. Nasir, G. D. F. Morales, D. Garcia-Soriano, N. Kourtellis, and M. Serafini. Partial key grouping: Load-balanced partitioning of distributed streams. *arXiv preprint arXiv:1510.07623*, 2015.

[6] M. A. U. Nasir, G. D. F. Morales, D. García-Soriano, N. Kourtellis, and M. Serafini. The power of both choices: Practical load balancing for distributed stream processing engines. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*, pages 137–148. IEEE, 2015.

[7] M. A. U. Nasir, G. D. F. Morales, N. Kourtellis, and M. Serafini. When two choices are not enough: Balancing at scale in distributed stream processing. In *Data Engineering (ICDE), 2016 IEEE 32nd International Conference on*, pages 589–600. IEEE, 2016.

[8] M. A. U. Nasir, F. Rahimian, and S. Girdzijauskas. Gossip-based partitioning and replication for online social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 33–42. IEEE, 2014.