Natali Ruchansky

Research statement

My research is centered around algorithmic aspects of data mining and machine learning for large datasets. I focus on integrating theoretical results with practical algorithms to tackle challenges that arise with real-world messy data, such as missing information, noise, and lack of structure.
I am particularly interested in drawing connections with problems in other disciplines and extracting insights from data that are meaningful both for layman and experts. My work has application across a wide range of areas including recommender systems, biology, neuroscience, social networks, urban informatics, and misinformation.

Previous and current research

Sparse data analysis

Data collected in the real world is often subject to resources or monetary constraints, resulting in datasets with unknown or missing values. Partially known datasets pose a challenge in applications such as recommender systems, biology, and urban informatics. One example occurred after the 2013 Boston Marathon when many runners could not finish the race, and recovering the times and qualification statuses of these runners required an estimation of missing values in data. My main branch of research is on developing efficient and accurate algorithms for tackling such problems. Particularly, I have focused on the problem of low-rank matrix completion which makes use of similarity within the data to estimate missing entries in a matrix.

Most real-world datasets do not satisfy the assumptions required by traditional matrix completion algorithms because the corresponding matrices are too sparse and not observed uniformly at random. As a result, the accuracy of the estimate is not guaranteed which limits its utility in practice. My work builds upon results from algebraic geometry to identify and overcome the cases in which traditional approaches are likely to fail. Namely, we propose a new framework for matrix completion that allows an analysis of the extent to which the input can be completed instead of blindly applying a completion algorithm[1]. Our key insight is to analyze the input and identify portions of the matrix that can be completed accurately even when the whole matrix cannot. We propose an efficient algorithm, called `completeID`, for identifying completable submatrices that is based on advances in graph mining. Our framework provides users with feedback on which parts of the estimate can be trusted, and has implications for the cold start problem.

Even with the ability to isolate the portions of the data that can be estimated accurately, data practitioners may want to estimate the remaining unknown entries. Often, these practitioners have the capability to add a small amount of observations to the partially-observed data. For example, Netflix can send a user survey to collect additional rating information. In this setting, a natural question is how much and which additional entries can be added to a non-completable matrix so that it becomes completable? To address this question, we built upon elegant theoretical results to transform the problem into an instance of infection prorogation on a graph[4]. By analyzing the infection, the algorithm identifies which entries need to be added for the partially-observed matrix to be completable. We showed that not only is the number of entries selected close to the lower bound, but the algorithm results in a more accurate estimate than existing work.

Structure in data

The second theme in my research is identifying and leveraging structure within data. In the matrix completion work discussed above, a major component in obtaining accurate solutions is the low-rank structure present in the data. While the low-rank assumption has proven to be a powerful mathematical assumption, it is often too simplistic to capture real world data. In the presence of low-rank submatrices, traditional matrix completion algorithms do not produce an accurate estimate. To overcome this we propose an algorithm that produces a more accurate completion by targeting low-rank submatrices in the input[5]. A key contribution in this work was the development of a simple algorithm for identifying low-rank submatrices in fully and partially known matrices.

Low-rank structure and (in)dependence are a common and natural way to capture the relationship among entities in a matrix (rows and columns). In graphs, the relationships among groups of proteins, users, routers are more difficult to capture. However, we observe that shortest paths between vertices can be seen as the analogy of rank structure in matrices. In fact, works in chemistry and mathematics have developed rigorous theory surrounding a quantity called the Wiener Index which evaluates as the sum of the shortest path between all pairs of vertices. We build upon this theory and leverage the shortest-path structure to gain insight into hidden relationships of vertices in large graphs. Specifically, we ask: *given a large graph and a few query vertices of interest, how can we learn about the relationships among query vertices?*[2]. We introduce the notion of the Minimum Wiener Connector as a subgraph that connects the query vertices and minimizes the Wiener Index. The constant-factor approximation algorithm we develop finds connectors that prove meaningful both in structure and context across many important applications such as social networks, biology, traffic, and more. For example, through experiments on a protein-protein interaction network, we found that the algorithm uncovers potential disease-associations of the input query vertices.

Future research

In my future research, I will continue to work towards developing reliable, accurate algorithms for tackling challenges that arise due to the messy nature of real world data. By making connections with problems in different disciplines, I hope to leverage elegant theoretical results to develop efficient algorithms that can be applied to practical domains.

In the near future of my postdoc, I am focusing on higher-order relationships in data. While matrix methods are extremely powerful, data gathered in the real world often contains more than two dimensions, hence it can be naturally modeled as a tensor. Even with only one additional dimension, most fundamental matrix problems already become much harder. Currently, I am working on developing an algorithm that can identify completable subtensors. Aside from providing confidence feedback on the estimate, an important consequence of this work is that it allows for a more targeted application of tensor completion algorithms, improving both accuracy and efficiency.

In another line of work, I am focusing on incorporating the temporal dimension into data analysis. Specifically, I have been working on capturing the temporal evolution of social interaction to target the fake news detection problem[6]. By leveraging the power of deep neural networks, traditional linguistic analysis can be combined with the time dimension to predict fake news as well as malicious users.

Finally, I am also working on higher-order analysis in graphs. Historically, graphs have been studied from an edge-centric point of view; however, real-world networks are composed of relationships that are more complex. In my work across two different projects, I am developing algorithms that can capture more complex connectivity patterns in graphs, as well as the property of vertices not being connected[3] – relationships that prove meaningful in areas such as neuroscience.

Publications

[1] D. Cheng, N. Ruchansky, and Y. Liu. Matrix completion with graphs: Identifying completable submatrices via edge connectivity. In *Under Submission*, 2017.

[2] N. Ruchansky, F. Bonchi, D. García-Soriano, F. Gullo, and N. Kourtellis. The minimum wiener connector problem. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1587–1602. ACM, 2015.

[3] N. Ruchansky, F. Bonchi, D. García-Soriano, F. Gullo, and N. Kourtellis. To be connected, or not to be connected: That is the minimum inefficiency subgraph problem. In *Under Submission*, 2017.

[4] N. Ruchansky, M. Crovella, and E. Terzi. Matrix completion with queries. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1025–1034. ACM, 2015.

[5] N. Ruchansky, M. Crovella, and E. Terzi. Targeted matrix completion. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2017.

[6] N. Ruchansky, S. Seo, and Y. Liu. Csi: A hybrid deep model for fake news. *arXiv preprint arXiv:1703.06959*, 2017.