Overview

A central challenge in the Information Age is to analyze the wealth of data generated in many domains, such as social networks, bioinformatics, and the Internet. Answers to relevant questions about such data often take the form of solutions to carefully formulated optimization problems. Examples of such questions include:

- In the context of a large-scale, dynamic communications network, such as the Internet, how badly is communication impaired upon the failure of elements in the network?
- Given the spread of a false news story on a social network such as Facebook, how could this misinformation be contained or limited effectively, while minimizing the impact on the social network?
- Given terabytes of genetic sequencing data[1], how can redundancies be eliminated to represent the data succinctly, while enabling the assembly of the entire, original genome?

Unfortunately, even under relatively simplistic modeling assumptions, many practical questions result in NP-hard optimization problems, which makes efficient, exact solutions unlikely. Even if a question is answerable in polynomial time, in many applications data has reached such proportions that a time complexity of $\Omega(n^2)$ is prohibitive, where n is the number of distinct data elements. Moreover, typical data sets are dynamic, so the solution of an algorithm may lose its relevancy even before the algorithm completes. As an example of dynamism, consider a wireless network where nodes constantly enter and leave the network.

Motivated by these challenges, my research centers around the design of **scalable and dynamic algorithms** that run on billion-scale datasets and retain a **performance guarantee** with respect to the optimal solution. A related theme is the study of **dynamic, succinct representations** of complex data. Developing effective and practical solutions requires a confluence of disciplines including approximation algorithms, graph theory, high performance computing, and computational biology.

Specifically, my research methodology comprises the following four steps.

1. **Accurate modeling of motivating applications** to capture the desired phenomena. For example, how should the diffusion of information in a social network be modeled? Such modeling is necessary to convert a vague question into a rigorous problem statement. Examples of modeling contributions on social networks include a the formulation of heterogeneous diffusion models [7], contributions to modeling of the time-scale of social diffusion [15], and a model that includes multiple products [18].
2. **Formulation of optimization problems** to answer the original motivating question. For example, for viral marketing on a social network, I formulated [11] an optimization problem that ensures a given level of exposure in the social network.
3. **Design of efficient algorithms and data structures** with theoretical guarantees on performance. Once the problem has been rigorously formulated, a provably scalable algorithm is formulated, and the loss in performance with respect to an optimal solution is quantifiable. Most of my works include theoretical results on worst-case time and space complexity.
4. **Empirical validation** of the proposed approaches in practical settings. For example, once a data structure for genome assembly was developed and theoretically analyzed, we evaluated how it performs on real sequencing data [4].

My research contributions encompass a broad range: from theoretical advances, such as my formulation [13] of the first polynomial-time approximations for weakly submodular maximization on the integer lattice; along with practical implementations, such as my implementation [4] of a dynamic data structure for use in genome assembly. These examples illustrate the end-to-end value of my approach.

[1] Modern sequencing technology produce billions of redundant, overlapping 100 to 150 base pair sequence reads.

Prior work: dynamic vulnerability assessments of complex networks An important measure of a large-scale communications network, such as the Internet, is how vulnerable the network is to attack or random failure. In order to quantify network robustness, a suitable metric as a proxy for communication is required; previous studies used *connectivity*, or the presence of a path between two nodes, as a proxy for communication.

My dissertation generalized the metric of connectivity to other measures, including *clustering* [10], *short cycles* [8], and *distance-based* [12, 9] measures of functionality. These metrics formed the basis of new optimization problems to measure network robustness. To approximate these vulnerability assessments, I proposed novel scalable and dynamic algorithms. The algorithms for the metrics of short cycles and paths are able to run on billion-scale networks and update their solutions in microseconds in response to changes in the network, and they maintain a competitive ratio with respect to the optimal solution. For the assessment based upon short paths, I proved that this ratio is optimal under plausible complexity-theoretic assumptions; that is, it is unlikely that an algorithm with a better worst-case guarantee exists, even for the case of a static network.

Prior work: other topics In the context of advertising on online social networks, I studied a new optimization problem, in which the level of total exposure in the network is ensured. For this problem, I developed the first scalable approximation algorithm. In addition, I proposed [7] a novel combination of heterogeneous diffusion models, for which I provided a scalable algorithm with performance guarantee, capable of running on billion-scale social networks. In addition, I formulated [18] scalable approximation algorithms for a new model of the influence maximization problem with multiple products. Other contributions on space-efficient and scalable algorithms include *e.g.* [14, 13, 2].

Future Directions

Optimization for machine learning

Convex optimization has a long history of study: any local minimum of a convex function is also a global minimum. However, important applications in machine learning do not exhibit convexity or its discrete analogue, submodularity. For example, **training deep neural networks is a non-convex optimization problem** that is not well understood. In the discrete setting, non-submodular optimization problems include misinformation containment discussed below and other data science applications such as video summarization. **My preliminary work** in this direction includes the development of a novel algorithm for **fast maximization of non-submodular functions on the integer lattice [13]**, with application to viral marketing with partial incentives. This algorithm has a constant performance ratio and exploits the non-submodularity of the objective to decrease its running time.

Further research plans in this direction include the following.

- *Anytime approximation algorithms.* Traditionally, an algorithm returns a (partially) feasible solution for a given problem instance; for an NP-hard optimization problem, one is often interested in worst-case and average-case bounds on runtime and solution quality.

  However, the amount of computational resources available to devote to a given problem instance may vary. In this case, it is of interest to develop algorithms whose **performance guarantee is a function of the amount of resources the algorithm has used thus far.** In this case, the user is free to terminate the algorithm at any time and obtain a solution whose quality in comparison with the optimal is bounded.

- *Learned approximation algorithms.* Many optimization problems, such as the influence maximization problem in viral marketing, have input that follows certain distributions. For example, social networks often follow a power-law degree distribution; such networks may be decomposed into pieces on which an exact solution is easier to find. Inspired by the use of machine learning techniques to speed up

traditional indexing data structures [6], **I plan to apply deep learning methods to reduce a problem instance into easier subinstances**, which may be solved exactly using integer programming techniques, approximated with traditional approximation algorithms, or further layers of learned models.

- *Scalable algorithms for non-submodular and non-convex optimization.* Stochastic gradient descent is used to train neural networks in practice; this tends to yield good results despite the fact that this is a non-convex optimization problem. This behavior is not well understood, and **there is a lack of theoretical justification or worst-case guarantees on neural network training. I plan to design and theoretically analyze fast algorithms for non-submodular and non-convex optimization**. Effective approaches would include the ability to handle matroid constraints and generalized cost constraints that arrive in applications such as video summarization and would provide fast algorithms with theoretical performance guarantees that hold when the objective is non-submodular or non-convex. Techniques for non-submodular and non-convex optimization can be bridged through the continuous extension of discrete functions.

Prospective funding for this work includes the Faculty Research Award of Google AI, the Amazon Research Award, and DARPA.

Information propagation on social networks

Misinformation may consist of a false or misleading news story shared in a social network such as Twitter; the objective is to slow down or contain the spreading of such information while minimizing the impact on the social network. There is a need for countermeasures in which the spread of misinformation in a social network is contained. For example, inoculating critical users before the misinformation reaches them may help contain or inhibit the spread. Another countermeasure would be to facilitate the spread of competing information. With collaborators, I have developed **preliminary measures and heuristics in this direction**, including early detection of misinformation through monitor placement [17] and a combat score [16] that measures the success of misinformation containment.

Future work on information propagation social networks includes the following.

- *Use real cascade data to approximate the influence function.* Much prior work studying problems concerning information propagation on social networks relies on simple, combinatorial models of propagation, such as the Independent Cascade model. However, these models do not explain real cascade data well. Therefore, **I plan to employ deep learning models to predict future cascades based upon past cascade data.**
- *Develop non-submodular approximation algorithms for election defense.* If the goal of the attacker is to influence the outcome of an election by targeting key districts, rather than simply maximizing the number of influenced people, the objective function is non-submodular. Hence, an effective defense strategy must account for this non-submodularity.
- *Study model-independent approaches for misinformation restraint.* Some of my preliminary results on misinformation restraint are discussed above. However, these methods rely on specific cascade models, such as Independent Cascade, that may be inaccurate. Hence, there is a need to **develop restraint approaches that work well across a variety of propagation models.** Also, social networks change on the same timescale as propagation, so it is important to develop dynamic solutions that can be efficiently updated upon network changes.

Prospective funding for this research includes the NSF program Resource Implementations for Data Intensive Research in the Social, Behavioral and Economic Sciences (RIDIR).

Summarization of biological data

Genomic databases continue to rapidly increase in size (the $100,000$ genomes project is projected to be more than 300 terabytes). This increased size creates new challenges: previously used algorithms and data structures for analysis of this data are no longer effective.

For example, the *de Bruijn graph* is a network commonly used for genome assembly from short-read data. Due to sequencing errors and ambiguities, there is a need for dynamic data structures capable of vertex / edge updates without rebuilding the entire data structure from scratch. We undertook a first step in this direction [4], in which **a practical, space-efficient, and dynamic de Bruijn graph data structure was provided**, with support for arbitrary edge insertions / deletions, and a small number of node insertions / deletions. However, when de Bruijn graphs are built for data from multiple biological species, it is natural to associate a subset of *colors* with each edge, to indicate which species the edge comes from. Therefore, the following specific goals are identified:

- *Develop a de Bruijn graph data structure that is both succinct and fully dynamic.* To accomplish this requires new implementations of data structures such as a dynamic wavelet tree over a string and dynamic bit vector representations. In addition to significantly improving the efficacy of genome assembly by yielding a dynamic de Bruijn graph representation close to the size of [3], these dynamic implementations will be useful in a variety of other contexts, such as dynamic compression algorithms and dynamic FM indexes [5].
- *Improve the succinct representation of color data in de Bruijn graph representations.* With collaborators, I developed a strong inapproximability result for efficient coloring [1], but open problems remain in succinct color storage, dynamism, and re-coloring algorithms. Results in this direction will allow better identification of genetic variation between and within species, with the potential to aid in the identification of genes responsible for inherited traits.

In addition, there is a need for **indexes of these genomic databases capable of supporting pattern-matching and search requests.** Currently, genomes are indexed by the FM index [5], which is closely related to the Burrows-Wheeler transform (BWT). In **preliminary work**, we developed [2] a **space-efficient algorithm** to quickly compute the BWT, designed to work best in the **context of repetitive genomic databases**. However, the BWT is only one ingredient of an efficient, large-scale index, and many algorithmic challenges remain before truly practical indexes are developed for these large databases. For example, space-efficient and scalable construction of the suffix array of a string is needed. Potential funding for this research includes initiatives of NIH Strategic Plan for Data Science, including Big Data to Knowledge (BD2K).

Career Plan

My career plan is to make significant contributions through the development of novel theoretical methods and practical implementations for problems arising from a variety of data science applications. As historically, a major emphasis will be realistic modeling that yields rigorous optimization problems whose solutions may be approximated with provable performance guarantees. The end-to-end focus of my methodology will ensure that, in addition to novel theoretical results, practical tools – such as a practical dynamic de Bruijn graph implementation or scalable implementations for misinformation containment – will benefit researchers from a variety of disciplines, such as bioinformatics or social sciences. As an academic advisor, I will adapt these resources into objectives suitable for graduate and undergraduate researchers as a component of their education.

References

[1] Bahar Alipanahi, Alan Kuhnle, and Christina Boucher. Recoloring the Colored de Bruijn Graph. In *Symposium on String Processing and Information Retrieval (SPIRE)*, 2018.

[2] Christina Boucher, Travis Gagie, Alan Kuhnle, and Giovanni Manzini. Prefix-Free Parsing for Building Big BWTs. In *18th International Workshop on Algorithms in Bioinformatics (WABI)*, 2018.

[3] Alexander Bowe, Taku Onodera, Kunihiko Sadakane, and Tetsuo Shibuya. Succinct de Bruijn Graphs. In *Workshop on Algorithms in Bioinformatics (WABI)*, pages 225–235, 2012.

[4] Victoria Crawford *, Alan Kuhnle *, Christina Boucher, Rayan Chikhi, and Travis Gagie. Practical Dynamic de Bruijn Graphs. *Bioinformatics*, 2018.

[5] Paolo Ferragina and Giovanni Manzini. Indexing Compressed Text. *Journal of the ACM*, 52(4):552–581, 2005.

[6] Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, and Neoklis Polyzotis. The Case for Learned Index Structures. In *SIGMOD*, 2018.

[7] A. Kuhnle, M.A. Alim, X. Li, H. Zhang, and M.T. Thai. Multiplex Influence Maximization in Online Social Networks With Heterogeneous Diffusion Models. *IEEE Transactions on Computational Social Systems*, 2018.

[8] A. Kuhnle, V.G. Crawford, and M.T. Thai. Scalable and Adaptive Algorithms for the Triangle Interdiction Problem on Billion-Scale Networks. In *IEEE International Conference on Data Mining (ICDM)*, 2017.

[9] Alan Kuhnle, Victoria G Crawford, and My T Thai. Network Resilience and the Length-Bounded Multicut Problem: Reaching the Dynamic Billion-Scale with Guarantees. *Proc. ACM Meas. Anal. Comput. Syst.*, 2(1), 2018.

[10] Alan Kuhnle, Nam P Nguyen, Thang N Dinh, and My T Thai. Vulnerability of clustering under node failure in complex networks. *Social Network Analysis and Mining*, 7(8), 2017.

[11] Alan Kuhnle, Tianyi Pan, Md Abdul Alim, and My T. Thai. Scalable Bicriteria Algorithms for the Threshold Activation Problem in Online Social Networks. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2017.

[12] Alan Kuhnle, Tianyi Pan, Victoria G Crawford, Md Abdul Alim, and My T Thai. Pseudo-Separation for Assessment of Structural Vulnerability of a Network. In *Proceedings of SIGMETRICS '17*, Urbana-Champaign, IL, 2017.

[13] Alan Kuhnle, J. David Smith, Victoria G. Crawford, and My T. Thai. Fast Maximization of Non-Submodular, Monotonic Functions on the Integer Lattice. In *International Conference on Machine Learning (ICML)*, 2018.

[14] S. Mishra, X. Li, A. Kuhnle, M.T. Thai, and J. Seo. Rate alteration attacks in smart grid. In *Proceedings - IEEE INFOCOM*, volume 26, pages 2353–2361, 2015.

[15] Tianyi Pan, Alan Kuhnle, Xiang Li, and My T. Thai. Popular Topics Spread Faster: New Dimension for Influence Propagation in Online Social Networks. In *International Conference on Data Mining (ICDM)*, pages 1–11, 2017.

[16] Huiling Zhang, Alan Kuhnle, J. David Smith, and My T. Thai. Restraining Misinformation and Pushing out the Truth. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018.

[17] Huiling Zhang, Alan Kuhnle, Huiyuan Zhang, and My T Thai. Detecting Misinformation in Online Social Networks Before it is Too Late. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2016.

[18] Huiyuan Zhang, Huiling Zhang, Alan Kuhnle, and My T Thai. Profit Maximization for Multiple Products in Online Social Networks. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2016.