Research Statement

Evimaria Terzi
Computer Science Department, Boston University

The proliferation of data produced by different applications has created ample opportunities for data-mining research. In this era of "big data", the goal of my research is to find efficient and principled methods for extracting useful knowledge from large data collections. I usually start with data-analysis problems that arise in real applications. Then, I focus on finding clean computational formulations to model these problems and also on designing efficient algorithms for solving them. In terms of problem formulations, I am interested in simply stated, yet technically challenging, questions. Technically, I am interested both in the design and analysis of the algorithms as well as their practical performance on real datasets.

My research agenda focuses on the development of principled and simple algorithms that are efficient, robust to noise and can handle both noisy and partially-observed data. Although at different times I am working on different application domains, I always stay faithful to this vision.

Below, I discuss my current and future application-specific research contributions as well as my current and ongoing work on cross-cutting research themes.

Application-specific research themes

Team formation in expertise management systems: The proliferation of expertise management systems (e.g., `LinkedIn`, `Upwork` etc.) has led to research on team-formation problems. Our work on team formation [11] was the first to address the problem of hiring a team of experts with different skills and good history of collaborating together. Ever since, there has been a lot of work in the data-mining community around this topic. Our more recent work on this topic focuses on: (a) the development of team-formation methods that have their roots in practical observations as obtained by social scientists [2], (b) the development of algorithms for identifying the team members that need to be hired/fired and outsourced by a company to achieve optimal cost for projects that arrive in an online fashion [1], and (c) problems related to the minimum alteration of existing, yet problematic, teams so that their operation becomes smoother [10].

My recent collaboration with BU Spark! and the Hariri Institute of Computing (HIC) has given my research on team formation an additional, more practical, dimension as we are applying our team-formation algorithms to the teams formed in student hackathons. At the moment, we apply our algorithms to hackathons in the Boston area, but we plan to license our software to hackathons throughout the country.

In the future, I will consider questions like: "Given team performance how can we evaluate the ability level of the each individual team member?" or "Can we predict the performance of a newly-formed teams?". The difficulty of these questions is due to the fact that the performance of a team is not a summation of its individual experts. Rather, it is defined by an *unknown* function, which will be part of the learning task.

Urban data analysis My previous work on unconventional types of recommender systems increased my interest in recommendation of routes in urban environments [6, 7] and subsequently the development of algorithmic techniques for analyzing data that capture urban activity. Our recent work [15] has developed new algorithmic tools for identifying popular routes that city dwellers take in a city. This helped us find city hotspots both for locals and tourists. Motivated by the

widespread use of ride-hailing platforms (e.g., Uber or Lyft), we have developed a framework for recommending to drivers strategies that maximize their expected earnings [3]. Our methods for this problem allow for significant improvement of the earnings of drivers when compared to naive drivers that are not strategic or simply follow price surge.

In the future, I will continue developing algorithmic techniques for analyzing data coming from different applications designed to improve different modes of peoples' lives.

Social networks: The analysis of social-network data has been a recurring theme in my research. Currently, I am interested in some recurring questions related to network sparsification [8] or the identification of functional communities [5]. More recently, motivated by the rise of echo-chambers I have worked on the problem of modifying the network so that such phenomena (and their effects) are minimized [13]. I expect that these problems will become vital in the future and my work on social networks will focus more and more on themes with such direct impact on society.

Cross-cutting research themes and long-term research vision

Apart from the domain-specific work, I also focus on more fundamental research questions that permeate existing data-mining approaches across multiple application domains.

Active data mining: In many applications the data come in the form of aggregates as massive scale makes fine-grained data collection and storage intractable. In other applications, we have access to only a small percentage of the data points. Most of the times, one needs to perform analysis of such incomplete or aggregated data either to perform prediction or other data-analysis tasks (e.g., matrix completion or clustering). Over the last years my research has focused on the development of techniques that appropriately query the data (e.g., in practice by asking specific feedback from users, or obtaining specific measurements) in order to better perform a specific data-mining task [4, 9, 12, 14, 16]. Although some of these papers are a bit older, this direction became quite dominant in my research over the last years.

The end goal of this research direction is to develop a generalized framework for what I call *active data mining*. This framework will allow us to strategically query unseen data with the goal to optimize for the data-mining task at hand (e.g., clustering, community detection, etc.).

Robust data mining: In our recent work on ride-hailing platforms [3], we developed a framework that enabled us to rigorously reason about and analyze the sensitivity of our results to perturbations in the input data, without ever performing these perturbations. Motivated by this, I plan to further investigate the application of similar ideas to other data-mining problems. On a high level, this research thrust will address long-standing questions in data mining wrt the impact of noise in the results. For this, we will build upon existing work on robust optimization, yet we will focus on the robust versions of data-mining problems, which have not been studied rigorously before.

Application of primal-dual algorithms in data-mining applications: A very recent and exciting direction to my research has been the application of primal-dual algorithms to a large number of data-mining problems. Although many of the algorithms we use in the data-mining community have their roots in some primal-dual algorithm, we – as a data-mining community – rarely go back to the original linear-program formulation of our problems. The goal of my future research in this domain is to do exactly that as we have ample preliminary evidence that shows that these algorithms can become very efficient and useful when they take into consideration the structure imposed by the practical problems we use them for.

References

[1] A. Anagnostopoulos, C. Castillo, A. Fazzone, S. Leonardi, and E. Terzi. Algorithms for hiring and outsourcing in the online labor market. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD*, pages 1109–1118, 2018.

[2] S. Bahargamam, B. Golshan, T. Lappas, and E. Terzi. A team-formation algorithm for faultline minimization. *Expert Systems with Applications*. To appear.

[3] H. A. Chaudhari, J. W. Byers, and E. Terzi. Putting data in the driver's seat: Optimizing earnings for on-demand ride-hailing. In *ACM International Conference on Web Search and Data Mining, WSDM*, pages 90–98, 2018.

[4] H. A. Chaudhari, M. Mathioudakis, and E. Terzi. Markov chain monitoring. In *SIAM International Conference on Data Mining, SDM*, pages 441–449, 2018.

[5] E. Galbrun, B. Golshan, A. Gionis, and E. Terzi. Finding low-tension communities. In *SIAM International Conference on Data Mining, SDM*, pages 336–344, 2017.

[6] E. Galbrun, K. Pelechrinis, and E. Terzi. Urban navigation beyond shortest route: The case of safe paths. *Inf. Syst.*, 57:160–171, 2016.

[7] A. Gionis, T. Lappas, K. Pelechrinis, and E. Terzi. Customized tour recommendations in urban areas. In *ACM International Conference on Web Search and Data Mining, WSDM*, pages 313–322, 2014.

[8] A. Gionis, P. Rozenshtein, N. Tatti, and E. Terzi. Community-aware network sparsification. In *SIAM International Conference on Data Mining, SDM*, pages 426–434, 2017.

[9] B. Golshan and E. Terzi. Unveiling variables in systems of linear equations. In *SIAM International Conference on Data Mining, SDM*, pages 145–153, 2014.

[10] B. Golshan and E. Terzi. Minimizing tension in teams. In *ACM on Conference on Information and Knowledge Management, CIKM*, pages 1707–1715, 2017.

[11] T. Lappas, K. Liu, and E. Terzi. Finding a team of experts in social networks. In *KDD*, pages 467–476, 2009.

[12] E. Malmi, A. Gionis, and E. Terzi. Active network alignment: A matching-based approach. In *ACM on Conference on Information and Knowledge Management, CIKM*, pages 1687–1696, 2017.

[13] A. Matakos, E. Terzi, and P. Tsaparas. Measuring and moderating opinion polarization in social networks. *Data Min. Knowl. Discov.*, 31(5):1480–1505, 2017.

[14] C. Mavroforakis, D. Erdös, M. Crovella, and E. Terzi. Active positive-definite matrix completion. In *SIAM International Conference on Data Mining, SDM*, pages 264–272, 2017.

[15] S. M. Nikolakaki, C. Mavroforakis, A. Ene, and E. Terzi. Mining tours and paths in activity networks. In *World Wide Web Conference on World Wide Web, WWW*, pages 459–468, 2018.

[16] N. Ruchansky, M. Crovella, and E. Terzi. Matrix completion with queries. In *ACM SIGKDD*, pages 1025–1034, 2015.