Research Statement

Dóra Erdős

Research philosophy

In the era of "Big Data" problems can only be solved by the **rigorous and methodological design of efficient algorithms**. This entails the use of good abstractions, clever data structures and exploiting the mathematics behind a context. This observation is the driving force of my research.

In my projects my research methodology always follows the same life cycle; (1.) I start with a real-life application, (2.) abstract and formalize the underlying problem, (3.) design efficient algorithms and (4.) test the algorithms on the real data. Ideally, the algorithms I find are general and can be applied to domains beyond the motivating application. My work, which I will describe below, in reconstructing graphs [2, 9], group centrality [7, 8, 10] and Boolean tensor factorization [4, 5] all share this life cycle.

Overall, in my research there is emphasis on **theoretical importance** and **practical relevance**. My research in pure mathematics [1] has deepened my skills in abstract modeling and methodological rigor. At the same time, my internship in industry (at American Express) strengthened my understanding of real world data mining problems.

My current research

Graph reconstruction. My work on graph reconstruction addresses the following question; *Given some aggregate information on the data can we reconstruct the original graph?*. This question is motivated by the increasing demand for privacy; often the original data are kept private and instead some aggregate information is made public. Common understanding is that such an aggregate is *"safe"* to release as long as the underlying data (or part of it) cannot be recovered from the given information. We were the first to formulate this problem for the neighborhood information of graphs [2, 9]; we show that even when the only information released about the graph is the number of common neighbors that two vertices share, it is possible to reconstruct the graph almost perfectly. Our algorithmic solution features novel technique involving a rounding heuristic that makes it possible to achieve the reconstruction.

The reconstruction of data from aggregate information opens the door to a number of interesting questions to answer. For example, the neighborhood information of a graph is described by the Gram matrix of its adjacency matrix. My current work in this area explores how well we can reconstruct the underlying data matrix if only a part of its Gram matrix is given. We investigate whether the Gram matrix can be completed, whether there exist a unique completion and whether we can find completions that admit some additional property that is known of the underlying data (e.g., Is the data integer valued? Does it have low rank? Does it maximize the determinant?). We also present an active version of this problem; in cases when the data cannot be reconstructed from the current partial knowledge of the aggregate, can we ask a small number of targeted queries so that reconstruction is possible?

Group centrality. Determining the importance of a single node in the network – often referred to as the centrality of the node – has lately received wide attention. However, in many applications, nodes act interdependently and hence *we need to be able to characterize the importance of a group of nodes in the graph.* Group centrality was first formally defined by Everett and Borgatti [11]. We are the first to pose it as a combinatorial optimization problem; *find a group of k nodes with largest centrality.* Of course the definition of centrality is application dependent. In our work on Filter Placement [8], we were the first to propose the problem of reducing information overload in a network by equipping some strategically chosen nodes with filtering capabilities. In our paper [7], we introduce the problem of Repetition-Aware Content Placement in navigational networks. The novelty of this work is that we introduce the concept of memory both in the navigational network as well as the objective function. Finally, an important piece of my work [10] develops a general framework to compute the group centrality of any *paths-based* centrality measure. As a byproduct of our research on group centrality I dis-

covered a divide-and-conquer approach [3] that speeds up the famous algorithm of Brandes [12] for betweenness centrality.

One direction that I am eager to pursue is that of centrality engineering; given a graph and a group of nodes, edit the graph (given some budget) so that the centrality of the group is increased/decreased. The instances of this problem again depend on the specific theme to which it is applied. For example, an airline has major hubs in only a few airports and it is not realistic to demand from them to establish more hubs. Hence, the airline can improve its position on the market only by adding or canceling flight routes. The above airline could decide to cancel a direct flight between two remote cities and instead increase the number of flights connecting its hubs, this way becoming a more popular choice for travelers. Another direction in which I am very interested is that of building a tool, Centrality as a Service (CaaS). For this tool I plan to develop a query language that enables the user to define any path-counting-type centrality measure (e.g., betweenness, closeness, degree centrality). The tool will run a generalized algorithm for computing paths-based centralities that is then adapted to the user-specified centrality notion with help of the query input. This research can both lead to technical discoveries (e.g., how to define the query language, the universal algorithm for centrality, large-scale implementation) as well as research on human factors (e.g., what is a good interface for users with regard to ease of use versus customizeability, running time versus accuracy of the computed centrality).

Boolean tensor factorization. I started working on Boolean tensor decomposition during my internship at Max Planck Institute and have been working on this topic ever since. The motivating application of our work was to automatically generate synonym phrases based on data obtained via open information extraction. The output of this work can then be used, for example, to better train text classifiers. In this application we represent noun-verb-noun phrases that appear in the data with help of a Boolean tensor. Our work [5] is aimed at finding patterns in this data and then using it to generate phrases with synonym meanings. We were the first to employ Boolean tensor factorization for this task. Tensor factorization is a more difficult task than work with matrices; due to the cubic (or higher order) size of the data, and the fact that it is proven that no best rank-k approximation exists. In our work we propose a novel algorithm for Boolean CP and Tucker decompositions [4] (an extended version of [4] is available on arXiv [6]). Most previous algorithms do decomposition by fixing factors in a greedy fashion by, for example, alternating least squares optimization [13] or mining association rules [14]. However, my contribution was an algorithm to find an initial CP-decomposition by way of creating an appropriate auxiliary graph and performing multiple random walks on this graph.

In many problems, data can be expressed by way of a matrix or higher order tensor. Often we need to detect some "*dense*" part of the data; this is the case, for example, in biclustering, subspace clustering, matrix/tensor factorization or community detection in graphs. Although, for many applications what comprises a good density measure is straightforward, in other cases it is not clear and only ad-hoc solutions exist. For example how does one define a useful density notion – keeping the specific application in mind – in heterogeneous networks, real valued data, data with random factors, or time stamped data where similar patterns occur at different times? My goal is to explore many possible themes and find some general best practice approaches for defining density when the data have some property.

Future work

As a faculty member, I plan to follow the same research approach; (1.) become motivated by a real-life problem, (2.) find and apply a solution to this problem, and (3.) explore the possibility of applying the techniques I design to problems coming from domains other than the motivating one. I hope to explore many new themes, especially in emerging areas (e.g., cloud computing, MOOCs) as well as interdisciplinary research. For example, my newest research topic is in the area of education; using students' grade data from two departments, we are working on the problem of simultaneously characterizing student and course types. We hope to apply this to recommend classes that best fit their type to students. Another direction I plan to explore is that of building universal tools that can be applied to multiple problems in the same abstraction domain. Motivated by my past work on group centrality, I will start by building a tool that explores Centrality as a Service.

Besides finding interesting research topics, I believe it is also important to build new collaborations among researchers with different backgrounds fostering creative combination of the participants' skills. For this, I will seek out collaborations with other CS scientists (working in areas such as networking, cloud computing, e-commerce) as well as non-CS faculty members (e.g., working in biology, sociology or physics departments). I believe that my ability to engage with new application domains and design practical yet theoretically founded algorithmic solutions will benefit both sides of such collaborations.

References

[1] D. Erdös, A. Frank, K. Kun, Sink-stable sets of digraphs, *SIAM J. of Discrete Mathematics (SIDMA)*, (2014), vol. 28, (2014), no. 4, 1651-1674.

[2] D. Erdös, R. Gemullla, E. Terzi, Reconstructing graphs from neighborhood data, *ACM Trans. on Knowledge Discovery from Data (TKDD)*, vol. 8, (2014), no. 4, 23

[3] D. Erdös, V. Ishakian, A. Bestavros, E. Terzi, A divide-and-conquer algorithm for betweenness centrality, *SIAM Data Mining (SDM)* , (2015)

[4] D. Erdös, P. Miettinen, Walk'n'Merge: A scalable algorithm for Boolean tensor factorization, *IEEE Int. Conf. on Data Mining (ICDM)*, (2013), 1037–1042.

[5] D. Erdös, P. Miettinen, Discovering facts with Boolean tensor Tucker decomposition, *ACM Int. Conf. on Information Knowledge Management (CIKM)*, (2013), 1569–1572.

[6] D. Erdös, P. Miettinen, Scalable Boolean tensor factorization using random walks, *arXiv:1310.4843*, (2013)

[7] D. Erdös, V. Ishakian, A. Bestavros, E. Terzi, Repetition-aware content placement in navigational networks. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2013), 820 - 828.

[8] D. Erdös, V. Ishakian, A. Lapets, E. Terzi, A. Bestavros, The filter-placement problem and its application to minimizing information multiplicity, *PVLDB*, vol. 5, (2012), no. 5, 418–429.

[9] D. Erdös, R. Gemulla, E. Terzi, Reconstructing graphs from neighborhood data, *IEEE Int. Conf. on Data Mining (ICDM)*, (2012), 231–240.

[10] V. Ishakian, D. Erdös, E. Terzi, A. Bestavros, A framework for the evaluation and management of network centrality, *SIAM Int. Conf. on Data Mining (SDM)*, (2012), 427–438

[11] M. G. Everett, S. P. Borgatti, The centrality of groups and classes, *The Journal of Mathematical Sociology*, vol. 23. (1999), no. 3, 181-201

[12] U. Brandes, A Faster Algorithm for Betweenness Centrality, *Journal of Mathematical Sociology*, vol. 25, (2001), no. 163, 163-177

[13] T. G. Kolda and B. W. Bader, Tensor decompositions and applications, *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[14] Pauli Miettinen, Boolean Tensor Factorizations. *IEEE Int. Conf. on Data Mining (ICDM)*, pp. 447–456, 2011.