Jelena Tešić: Research Impact and Plan

Unstructured data encompasses diverse data types and formats, such as video data, electronic health data records, surveys, experimental readings, sensor data, social platform content data, genomics data, and purchase reviews. The arching quest of my research is to **identify** analytics tasks where unstructured data or task at hand do not adhere to underlying assumptions of state-of-art algorithms and to **provide** efficient, effective, intuitive, and responsible algorithms for addressing the challenges of extensive unstructured data collections in the wild. The research impact and plan Dr. Tešić lead in the Data Lab @ TXST (DataLab12.github.io) is grouped by the tasks it solves in the following sections.

1 Scaling Signed Graph Tasks through Fundamental Cycle Basis

In network science, signed graph representation of relations provides more information than unsigned graph networks. Spectral pollution, computational complexity, inherent bias, and training requirements are the main hurdles that have slowed the adoption of the signed network as data representation and signed graph algorithms for modern data analysis. This project drew inspiration from mathematical sociology and focused on balance theory to solve the signed network tasks at scale. First, the research has expanded the balance theory to signed social network graph analysis. Dr. Tešić w Dr. Rusnak proposed a frustration cloud view of the signed graph where the vertices and edges are quantified in frustration cloud statistics and validated this new social network graphs analysis approach for multiple real networks [31] [pdf] Dr. Tešić initialized the development of an algorithm to efficiently compute the fundamental cycle basis in large, unstructured graphs. This collaborative approach does not require training data and avoids spectral pollution to scale the frustration cloud computation [9]. There has been no benchmark to evaluate state-of-the-art tasks on signed graphs derived from real networks. Dr. Tešić has led the effort to compare and contrast state-of-art community discovery on actual signed graphs in [16] [pdf]. Methods are compared assumption-free in terms of efficacy, efficiency, scalability, and reproducibility of existing methods. As a follow-up, the team has proposed the frustration cloud-based approach for cluster boosting for high modular signed graphs [17] [pdf]. Two master and five undergraduate students contributed to this research project.

The research plan is to use a fundamental cycle basis to scale the solution of other NP-hard tasks to large signed networks using the efficient fundamental cycle computation approach [25] [pdf] and propose algorithms to solve the task of computing frustration and the balanced state of the graph at scale [27] [pdf] and the task of finding the largest balanced subgraph in any network [26] [pdf] for graphs with millions of nodes and edges. The project has evolved into a significant part of the Ph.D. thesis work. Dr. Tešić and her team are working on the scalable community discovery algorithm, extending the analysis using a fundamental cycle basis to the recommendation and anomaly detection tasks and applying the algorithms to sensor and agent networks and signed gene networks.

2 Modeling Social Network Relations

The multifaceted interconnectivity of users and content on Twitter through user connections, replies, quotes, hashtags, and shared content makes it an exciting medium for research on the effectiveness of the representation and methods used. The project has introduced a scalable end-to-end Twitter network data management pipeline that gathers, stores, and models rich relationships from Twitter networks [14] [pdf]. The research work compared and contrasted the analysis results of millions of Twitter data using multiple graph construction processing approaches [15] [pdf]. How well the tweet content can be classified based on modeling relationships from interactions alone, and how well can community classification predict the label of the content? The community-based modeling (tweet is classified not on the content but on the retweets, replies, quotes, hashtags, and the author) yields precision, recall, and accuracy comparable to lexical classifiers [24] [pdf]. The project proposes new multi-modal approaches that consistently deliver the most robust outcomes and exhibit the highest performance measures for network graphs constructed based on Twitter interactions related to the COVID-19 pandemic [1] [pdf] and [24] [pdf]. One undergraduate, one master, and one Ph.D. student participated in the research work for this research project. The next step is to extend relationship modeling to signed and weighted graphs and to explore network representations to provide an assumption-free and bias-free baseline for evaluating graphical neural network performance for network science tasks.

3 Identifying Small Objects in Highly Variable Overhead Videos

The high variability of content in the overhead imagery stems from the terrestrial region captured, the high variability of acquisition conditions, and the number and size of objects in aerial imagery that are very different than in the consumer data. State-of-art fails due to the high variability of the domain and the low availability of training data. The project has proposed multiple domain adaptation approaches to alleviate the degradation of object identification in previously unseen overhead datasets with significant domain gaps and dominant small objects [4] [pdf] and [28] [pdf]. The project has proposed new algorithms for overhead videos to detect anomalous activities [8, 32]. Five undergraduates, two masters, and three Ph.D. students participated in this project sponsored by NAVAIR. The project evolved in the Ph.D. thesis work, and the team proposed multiple innovative and efficient contrastive learning algorithms to improve object classification in previously unseen highly variable overhead datasets [5] [pdf] and [7] [pdf]. The team is now introducing progressive domain adaptation to produce domain-invariant features across aerial datasets using local and global components for domain adaptation and object classification for the task [6] [pdf] and [20] [pdf]. The study will focus on improving scene understanding in highly variable videos.

4 Deep Descriptor Database Indexing Search and Retrieval

Searching for unseen objects in extensive visual archives is challenging, demanding efficient indexing methods supporting meaningful similarity retrievals. Together with the Ph.D. student, Dr. Tešić has introduced new indexing and search algorithm for deep descriptor databases that have up to four times lower memory usage and higher effectiveness than state-of-art [21] [pdf] and [22] [pdf] on millions of deep descriptors. The study is building upon the approach for crowdsensing application [20] [pdf] and plans to improve the indexing footprint while keeping the search effectiveness by proposing a new stratified graph approach [23] [pdf]. Current research focuses on multi-modal indexing extension, combining feature indexing with GIS attribute, spatial, and shape indexing.

5 Semantic Segmentation Task in The Wild

Semantic image segmentation is the task of assigning a label to each pixel in an image. Dr. Teši'c's contribution to semantic segmentation focuses on pavement distress detection for transportation and road maintenance and NASA spaceflight sample images. The project has quantified the main influencing factors that affect the performance of deep learning models in pavement distress detection pipelines and proposed a semantic segmentation algorithm that significantly improves the accuracy of localizing pavement cracks [10] [pdf]. Current research explores the domain adaptation approaches from Section 3 to improve the algorithmic performance for specialized fields and how pixel-wise segmentation can improve the explainability of the model. In parallel, the project focuses on a semantic segmentation pipeline to automatically classify Bacterial Adhesion and Corrosion from images obtained in the NASA SpaceX-21 experiment. The inability to control microbial biofilms during spaceflight poses a severe health risk to astronauts. Automatically identifying corrosion in tens of thousands of images of samples flown into space will help researchers streamline the samples' data collection and help them reach conclusions faster on what countermeasures will work in space. Two undergraduates, two masters, and one Ph.D. student are actively involved in this project.

6 Predictive Modeling of Noisy Tabular Data

Tabular data in the wild are difficult to model due to the uneven distribution of attributes, missing, overlapping, noisy values, a mix of categorical and numerical data attributes, data imbalance, and a long tail of sparse values. The work has developed an intentional data science pipeline that can automatically uncover important attributes, reduce feature space, and model prediction in a robust manner from multi-source tabular data in [12]. Dr. Tešić's team has designed the multi-feature importance analysis algorithm and applied it to large-scale analysis of public data from the National Center for Education Statistics (NCES) to provide data-driven insights into teacher attrition challenges. The study discovered that the race and sex of the principal, the type of school, and the school's location impact teacher retention rates the most and that modeling historical data resulted in a predicted attrition rate of over 10%, aligning closely with the current prevalent attrition rates in the USA [13] [pdf]. The project has developed an interpretable data-driven scoring fusion to discover the most critical factors from an extensive collection of heterogeneous public data sources on learning loss during the COVID-19 pandemic in Tex s public schools. Our robust approach found that the number of students on school campuses in the Fall of 2020 and in the Spring of 2021 was the most resilient and most impactful predictor of how the students would perform on the standardized test in mathematics and reading in the Spring of 2021 in Texas [18] [pdf]. The work introduced new cascade enhancement to ensure effectiveness and the prediction coverage of our modeling pipeline to predict long COVID in N3C data [19] [pdf]. The project resulted in one honors thesis [3], one master thesis [12], and as a seed to a Ph.D. thesis work [19] [pdf]. Current work focuses on improving the loss function in the radiant boosting approach using contextual contrastive learning to handle the sparsity, relevance, and hierarchical values in the tabular N3C data more effectively. The work has been extended to include temporal modeling and the latest educational data for multi-year learning recovery predictions.

7 Summary

Algorithms for unstructured data are driven by the data type, format, acquisition, size, and intended task. My research focuses on providing new algorithms based on mathematics, computer science, and statistics theory while considering the specifications (efficiency, scalability, usability, interpretability) of the task at hand, data characteristics, and domain applicability. To this end, Dr. Tešić has collaborated with other Labs at our department and proposed data-driven solutions for their specific challenges [11, 29, 2, 30]. Dr. Tešić is currently looking into developing implicit neural representations to encode large climate models. My plan for Data Lab is to continue to invent new algorithms and methods for solving challenging analytics tasks stemming from the nature and size of unstructured data. * *author* - names of the the Data

Lab students are in *italic*

References Cited

[1] *Andrew Magill*, *Lia Nogueira de Moura*, *Maria Tomasso*, *Mirna Elizondo*, and **Jelena Tešić**. Enriching content analysis of tweets using community discovery graph analysis. In *Proceedings of the MediaEval 2020 Workshop*, volume 2882, 2020.

[2] *Blake W Ford*, Apan Qasem, **Jelena Tešić**, and Ziliang Zong. Migrating software from x86 to arm architecture: An instruction prediction approach. In *2021 IEEE International Conference on Networking, Architecture and Storage (NAS)*, pages 1–6, 2021.

[3] *Daniel Payan*. Ranking resilience attributes for texas public school districts. Honor's thesis, TXST, 2023. Advisor: **Jelena Tešić**.

[4] *David Heyse*, *Nicholas Warren*, and **Jelena Tešić**. Identifying maritime vessels at multiple levels of descriptions using deep features. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pages 423 – 431. SPIE, 2019.

[5] *Debojyoti Biswas*, *MMM Rahman*, Ziliang Zong, and **Jelena Tešić**. Improving the energy efficiency of real-time dnn object detection via compression, transfer learning, and scale prediction. In *The IEEE 16th International Conference on Networking, Architecture, and Storage (NAS 2022)*, September 2022.

[6] *Debojyoti Biswas* and **Jelena Tešić**. Domain adaptation with contrastive learning for object detection in satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing **(under review)***, 2023.

[7] *Debojyoti* Biswas and Jelena Tešić. Progressive domain adaptation with contrastive learning for object detection in the satellite imagery. In *arXiv:2209.02564*, 2023.

[8] *George E. Strauch*, *Jiajian (Jax) Lin*, and **Jelena Tešić**. Overhead projection approach for multi-camera vessel activity recognition. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5626–5632, 2021.

[9] *Ghadeer Alabandi*, **Jelena Tešić**, Lucas Rusnak, and Martin Burtscher. Discovering and balancing fundamental cycles in large signed graphs. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '21, New York, NY, USA, 2021. Association for Computing Machinery.

[10] *Haitao Gong*, **Jelena Tešić**, *Jueqiang Tao*, Xiaohua Luo, and Feng Wang. Automated pavement crack detection with deep learning methods: What are the main factors and how to improve the performance? *Transportation Research Record*, page 03611981231161358, 2023.

[11] *Hanie Samimi*, **Jelena Tešić**, and Anne Hee Hiong Ngu. Patient-centric data integration for improved diagnosis and risk prediction. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, pages 185–195, Cham, 2019. Springer International Publishing.

[12] *June Yu*. Gradient boosting public data modeling for the policy planning in education. Master's thesis, TXST, 2022. Advisor: **Jelena Tešić**.

[13] *June Yu*, Li Feng, and **Jelena Tešić**. Mitigating u.s. public school teacher attrition crisis: A data science approach. *Information Processing & Management **(under review)***, 2023.

[14] *Lia Nogueira de Moura*. Social network analysis at scale: Graph-based analysis of twitter trends and communities. Master's thesis, TXST, 2020. Advisor: **Jelena Tešić**.

[15] *Lia Nogueira de Moura* and **Jelena Tešić**. pytwanalysis: Twitter data management and analysis at scale. In *2021 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2021.

[16] *Maria Tomasso*, Lucas Rusnak, and **Jelena Tešić**. Advances in scaling community discovery methods for signed graph networks. *Journal of Complex Networks*, 10(3), 06 2022.

[17] *Maria Tomasso*, Lucas Rusnak, and **Jelena Tešić**. Cluster boosting and data discovery in social networks. In *Proceedings of the 37th ACM/SIGAPP Symposium On Applied Computing (SAC)*, 2022.

[18] *Mirna Elizondo*, *June Yu*, *Daniel Payan*, Li Feng, and **Jelena Tešić**. Data driven analysis of intervention effectiveness for covid learning loss in texas public schools. ***(Under Submission)***, 2023.

[19] *Mirna Elizondo*, Rasim Musal, *June Yu*, and **Jelena Tešić** on behalf of N3C. Long covid challenge: Predictive modeling of noisy clinical tabular data. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, 2023.

[20] *MMM Rahman*, *Debojyoti Biswas*, and **Jelena Tešić**. Evirec: Efficient visual indexing and retrieval for edge crowd-sensing. In ***Submitted to a Conference***, 2023.

[21] *MMM Rahman* and **Jelena Tešić**. Evaluating hybrid approximate nearest neighbor indexing and search (hannis) for high-dimensional image feature search. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 6802–6804, 2022.

[22] *MMM Rahman* and **Jelena Tešić**. Hybrid approximate nearest neighbor indexing and search (hannis) for large descriptor databases. In *2022 IEEE International Conference on Big Data*, pages 3895–3902, 2022.

[23] *MMM Rahman* and **Jelena Tešić**. Stratified graph indexing for efficient search in deep descriptor databases. In ***Submitted to a Conference***, 2023.

[24] *Muhieddine Shebaro*, *Lia Nogueira de Moura*, and **Jelena Tešić**. Improving association discovery throughmultiview analysis of social networks. *Social Network Analysis and Mining **(under review)***, 2023.

[25] *Muhieddine Shebaro* and **Jelena Tešić**. Identifying stable states of large signed graphs. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)*, 2023.

[26] *Muhieddine Shebaro* and **Jelena Tešić**. Abcd: Algorithm for balanced component discovery in signed networks. In ***Submitted to a Conference***, 2023.

[27] *Muhieddine Shebaro* and **Jelena Tešić**. Scaling frustration index and corresponding balanced state discovery for real signed graphs. In ***Submitted to a Conference***, 2023.

[28] *Nicholas Warren*, *Ben Garrard*, Elliot Staudt, and **Jelena Tešić**. Transfer learning of deep neural networks for visual collaborative maritime asset identification. In *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, pages 246–255, Oct 2018.

[29] *Noah Dunstatter*, *Alireza Tahsini*, Mina Guirguis, and **Jelena Tešić**. Solving cyber alert allocation markov games with deep reinforcement learning. In Tansu Alpcan, Yevgeniy Vorobeychik, John S. Baras, and György Dán, editors, *Decision and Game Theory for Security*, pages 164–183, Cham, 2019. Springer International Publishing.

[30] *Taylor Mauldin*, Anne H. Ngu, Vangelis Metsis, Marc E. Canby, and **Jelena Tešić**. Experimentation and analysis of ensemble deep learning in iot applications. *2019 VLDB DMAH*, 5(1):133–149, 2019.

[31] Lucas Rusnak and **Jelena Tešić**. Characterizing attitudinal network graphs through frustration cloud. *Data Mining and Knowledge Discovery*, 6, November 2021.

[32] **Jelena Tešić**, Dan Tamir, S. Neumann, Naphtali Rishe, and Abe Kandel. Computing with words in maritime piracy and attack detection systems. In Dylan D. Schmorrow and Cali M. Fidopiastis, editors, *Augmented Cognition. Human Cognition and Behavior*, pages 434–444, Cham, 2020. Springer International Publishing.