Socially Responsible

Data Science Day 2019

Friday May 10, 2019

Schedule

09:00 - 09:05 Welcome

09:05 - 10:30 Session on Responsible Data Science - *Moderator: Abel Rodriguez, Department of Statistics, UCSC*

Responsible Data Visualization - *Angus Forbes, Department Computational Media Department, UCSC*

The Professor and the Dashboard: A Cautious Approach to Classroom Data Analytics - *Jody Greene, Associate Vice Provost for Teaching and Learning; Director, Center for Innovations in Teaching and Learning; Department of Literature, UCSC*

Situated, Collaborative Modeling: Critical Participatory Data Science in Rural Zimbawe with the Muonde Trust -
Melissa Viola Eitzel Solera, Science and Justice Research Center, UCSC

Fair Algorithms - *Yang Liu, Department of Computer Science and Engineering, UCSC*

Data Privacy: From Theory to Practice - *Abhradeep Guha Thakurta, Department of Computer Science and Engineering, UCSC*

10:30 - 10:45 Break

10:45 - 11:45 Keynote address: Ethical Considerations in Data Science -
Mehran Sahami, Department of Computer Science, Stanford University

11:30 - 12:00 Poster Highlights

12:00 - 01:15 Lunch

01:15 - 02:30 Data Science for Social Good Session - *Moderator: Jim Whitehead, Department of Computational Media, UCSC*

Data Dividends?: Rethinking work and the commons in the era of big data - *Chris Benner, Dorothy E. Everett Chair in Global Information and Social Entrepreneurship; Director, Everett Program for Technology and Social Change; Director, Santa Cruz Institute for Social Transformation; Departments of Environmental Studies and Sociology*

Causal Inference Without an Experiment - *Carlos Dobkin, Department of Economics, UCSC and National Bureau of Economic Research*

Modeling for Seasonal Marked Point Processes: An Analysis of Evolving Hurricane Occurrences - *Athanasios Kottas, Department of Statistics, UCSC*

Using Big Data to Improve Students' Educational Outcomes in the Silicon Valley - *Rebecca London, Department of Sociology, UCSC*

Real-world benefits of Machine Learning in healthcare - *Narges Norouzi, Department of Computer Science and Engineering, UCSC*

02:30 - 02:45 Closing remarks

02:45 - 04:00 Poster session

1. Query Rewriting for Templated SRL Languages
2. When to Accept the Black Box
3. Interactive Story Generation
4. IGM-Vis: Analyzing Intergalactic and Circumgalactic Medium Absorption Using Quasar Sightlines in a Cosmic Web Context
5. BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2
6. Interactive Canvas Style Transfer
7. Aligning Product Categories using Anchor Products
8. Deep-Packet Analytics of SCADA Network Traffic Data
9. Art I Don't Like
10. Learning on Label Scarce Data
11. The power of pivoting in counting cliques
12. A self-exciting point process based on non-parametric Bayesian approach: modeling the pattern of occurrences of earthquakes
13. Multi-Scale Shotgun Stochastic Search for Large Spatial Datasets
14. Classification Cube
15. Efficiently Counting All 5-vertex Subgraph Orbits for Vertices
16. Entity Resolution on Online Games using Collective Social Behavior
17. Understanding the Factors that Affect Wellbeing after Intervention
18. Can Neural Generators for Dialogue Learn Sentence Planning and Discourse Structuring?
19. Comparing weight learning techniques of Probabilistic Template Languages
20. Global Edge Data Management
21. Bayesian Mixed Effect Sparse Tensor Response Regression Model with Joint Estimation of Activation and Connectivity
22. Lifted Hinge-Loss Markov Random Fields
23. Using Expression-based Variant Impact Phenotyping to Characterize Cancer Variants
24. MOBA Item Recommendation
25. Petfinder: Predict a Pet's Adoption Speed
26. Signal Processing of EEG data for Deep Neural Networks
27. Improving Tsunami Early Warning with Deep Learning
28. CruzAffect: a feature-rich approach to characterize happiness
29. NeuroCave
30. The Game of Protocols: Using Reinforcement Learning to Improve the Performance of Channel Access Schemes

Responsible Data Science Session

Responsible Data Visualization

Angus Forbes, Computational Media Department, UC Santa Cruz

Data visualization tools aim to help researchers find patterns and gain insight into complex datasets. However, visualization techniques are effective only if the data sources are reliable and if the models used to analyze the data are appropriate. Promoting transparency through emphasizing the provenance of data and the contexts in which the data is situated is an increasingly important component of contemporary data visualization. Creating effective representations and interactions is both an art and a science, and visualization benefits from an iterative design process which helps to clearly identify analysis tasks and encourage exploration while mitigating bias. This talk looks at recent projects from the UCSC Creative Coding Lab and discusses human-centered data visualization strategies.

The Professor and the Dashboard: A Cautious Approach to Classroom Data Analytics

Jody Greene, Associate Vice Provost for Teaching and Learning; Director, Center for Innovations in Teaching and Learning; Professor of Literature, UC Santa Cruz

New data analytics tools are being pioneered in the UC system and around the country to track and support student success at the classroom level. Some of these tools are already available to advisors and will soon be available to instructors at UCSC. What approach should the campus take to preparing the faculty to use data analytics, including predictive analytics, at the classroom level? What kinds of data should instructors have access to? And how can we support faculty to be informed and responsible users of this data?

Situated, Collaborative Modeling: Critical Participatory Data Science in Rural Zimbabwe with the Muonde Trust

Melissa Viola Eitzel Solera, Science and Justice Research Center, UC Santa Cruz

As a practicing modeler by training and experience, I have encountered many issues concerning the way data are handled, how models are made, and, in particular, how truth claims emerge from modeling. "Big data" are increasingly discussed in academic venues and elsewhere as an unquestioned good. However, there is dangerous potential for models to marginalize people based on biased or inappropriate data, affording them no recourse to those same tools to defend themselves. Therefore, modeling needs to be practiced more critically and less automatically. I suggest practices grounded in Haraway's "Situated Knowledge," requiring more descriptive methods and ways to model responsibly and collaboratively, and then apply these principles to my collaborative modeling efforts with the Muonde Trust in Mazvihwa Communal Area, Zimbabwe. We created an agent-based model in NetLogo to represent land-use decisions and other management interventions in Mazvihwa's agro-pastoral system, investigating the feedbacks and tradeoffs in land allocation to arable production versus woodland grazing area and the corresponding impact on livestock populations. We held workshops with community researchers, farmers, and leaders demonstrating and discussing the model at several points, which both led to alterations of the model and of land-use policy in the area. Local authorities are now in conversation with Muonde's researchers about ways to re-cultivate fallow fields rather than converting woodland to new arable production. Our modeling and community engagement process exemplifies many of the suggested situated modeling practices, and I will evaluate the success of critical participatory data science as a framing for this collaborative project.

Fair Algorithms

Yang Liu, Computer Science and Engineering, UC Santa Cruz

Machine learning (ML) is increasingly used in domains that have a profound effect on people's opportunities and well-being, including healthcare, law enforcement and consumer finance. A ML-trained system may appear to be fair without human intervention, this talk argues that this is far from being the case. A typical ML system consists of a pipeline of the following major components: data collection, model training and model deployment, and I’d like to give a high-level overview of the potential biases or discriminations that may arise in each of above components. This talk will also raise awareness of the following questions: 1) How to guarantee the quality of data collected from potentially careless and even malicious human agents, instead of treating the data as if they were clean and representative (blind trust)? 2) How to build ML methods that are robust despite noise in the data (bias in training data)? 3) How to guarantee fair and transparent treatment of people when ML models are deployed (bias in model and algorithm)

Data Privacy: From Theory to Practice

Abhradeep Guha Thakurta, Computer Science and Engineering, UC Santa Cruz

In this short talk I will give a brief overview of differential privacy, a widely used notion of statistical data privacy, and trace its journey from being a notion in theoretical computer science to large scale deployments in both industry and the government.

Keynote: Ethical Considerations in Data Science

Mehran Sahami, Associate Chair for Education and Director of Educational Affairs, Computer Science, Stanford University

Data science has tremendous potential to help us better understand a variety of domains and build tools for automated decision-making. Such tools carry the promise of more accurate predictions, greater insights, and much higher efficiency and throughput than might be achievable without them. However, data science also has the potential to lead to outcomes that reinforce biases, disproportionately impact particular subpopulations, and violate notions of privacy. In this talk I examine some of the promise and perils that arise from work in data science. I consider specific examples that allow us to take deep dives into ethical issues, such as algorithmic fairness and data privacy, to understand both the technical issues and competing value trade-offs at stake. My goal is not to find one "right answer," which I argue often does not even exist, but rather to help data scientists further appreciate the ethical considerations and value-laden implications of their work.

Data Science for Social Good Session

Data Dividends?: Rethinking Work and the Commons in the Era of Big Data

Chris Benner, Dorothy E. Everett Chair in Global Information and Social Entrepreneurship; Director, Everett Program for Technology and Social Change; Director, Santa Cruz Institute for Social Transformation; Professor, Environmental Studies and Sociology, UC Santa Cruz

In his state of the state presentation in January of this year, California Governor Newsom raised the idea that California should implement a data dividend, in which consumers would be paid for all the data they provide that companies monetize. In this talk, Dr. Benner will discuss this idea, along with his current work exploring the idea of a Universal Technology Dividend. He will explore questions related to the common-property characteristics of technology and innovation, the monopolistic characteristics of information markets, and the need to rethink how we define work in contemporary labor markets.

Causal Inference Without an Experiment

Carlos Dobkin, Economics Department, UC Santa Cruz and National Bureau of Economic Research

Unbiased estimates of treatments effects are critical for forming effective government policy. The Randomized Controlled Trial (RCT) is the gold standard approach to generating causal estimates. Unfortunately for many treatments of interest ethical or practical constraints make it impossible to implement an RCT. In some settings a Regression Discontinuity Design can be used to estimate the causal effect of a treatment. RDDs leverage rules that assign treatment eligibility based on the level of a variable. The sharp change in the probability of treatment induced by the rule make it possible to estimate causal effects as diverse as the effects of health insurance or alcohol consumption.

Modeling for Seasonal Marked Point Processes: An Analysis of Evolving Hurricane Occurrences

Athanasios Kottas, Statistics Department, UC Santa Cruz

Seasonal point processes refer to stochastic models for random events which are only observed in a given season. We present a Bayesian nonparametric modeling approach to study the dynamic evolution of a seasonal marked point process intensity. The motivating application involves the analysis of hurricane landfalls along the U.S. Gulf and Atlantic coasts from 1900 to 2010, for which the focus will be on the evolution of the intensity for the process of hurricane landfall occurrences, and the respective maximum wind speed and associated damages.

Using Big Data to Improve Students' Educational Outcomes in the Silicon Valley

Rebecca London, Sociology Department, UC Santa Cruz

Harnessing existing public agency data collections to improve the outcomes of children and youth seems like a simple and straightforward use of resources. However, there are legal, technical, and ethical barriers to cross-sector data sharing, impeding progress in this area. This presentation will explore the Silicon Valley Regional Data Trust, a data-sharing collaborative between education, juvenile probation, child welfare, and behavioral health agencies. It will highlight the ways that big data can be harnessed for social good, and also the challenges to overcome in creating systemic and ethical change to everyday data practices in youth-serving organizations.

Real-world benefits of Machine Learning in Healthcare

Narges Norouzi, Computer Science and Engineering, UC Santa Cruz

This talk will be a quick survey on all of our efforts in the Applied Machine Learning Lab in the Computer Science and Engineering department of the University of Santa Cruz. We will particularly discuss state-of-the-art machine learning healthcare applications and our efforts in those areas, including different data sources we are using and how data drives our investigations, ethics of using algorithms in healthcare, and future applications and direction of the lab.

Poster Session

1. Query Rewriting for Templated SRL Languages

Eriq Augustine, Theodoros Rekatsinas, Lise Getoor

Statistical Relational Learning (SRL) methods unify the complexity of structured models with the flexibility of probabilistic reasoning to produce accurate and robust models. Rule templating languages like Markov Logic Networks (MLN) and Probabilistic Soft Logic (PSL) have emerged as the predominant method of constructing SRL models. These languages are particularly well-suited to constructing models over richly structured domains, because they provide a logical formalism for describing entities and their relationships, and a compact mechanism for describing the parameters of the underlying model. However these models suffer from scalability issues, particularly when instantiating the full ground graphical model. In this poster, we create a method of speeding up this instantiation process by rewriting relational queries.

---

2. When to Accept the Black Box

Brian Blanchette

"From social media to search engines, people form beliefs from increasingly filtered information sources. However, the inner-workings of these sources are often unclear, which obscures subjects ability to judge the epistemic value of this information. As with many rapidly advancing technologies, issues of accountability and explicability have become a central worry of some regarding the opacity of certain applications of artificial intelligence. In this paper, we outline major types of machine learning technology, their respective uses, and challenges involved with each method. Importantly, we show that not only should we accept a so-called ""black box"" at times, but that we have no choice but to embrace a certain level of epistemic uncertainty when advancing technology.One key facet of living in our informationally informed society utilizing the product of newly created machine learning technologies. Machine learning uses algorithms and statistical models to perform a specific task without explicit instructions, relying on patterns and influence. Modern integration of machine learning techniques purport to answer our queries and suggest our future actions without direct instruction or detailed foraging.How can we know if a machine learning informed result is trustworthy? One way to know is to learn how the algorithm is trustworthy by coming to understand the inner-workings of each algorithmic model (Humphreys 2004). However, typical users cannot come to understand the foundational truths of machine learning algorithms because of issues of complexity and corporate secrecy which permeates through hardware, software, and algorithmic creation. Continuing elucidation of this problem by epistemologists and computer scientists, some have attacked information technology on the basis that the opacity of the algorithmic inner-workings may prevent users from critically assessing the process and criteria used in an algorithmic contemplation that produce a given output, in some cases, disallowing some users from fulfilling their epistemic responsibilities (Simon Ethics and Information Technology, 12(4), 343–355 2010). My framework considers several reasons that pave the pathway to acquiring epistemically responsible knowledge without becoming a PhD student in computer science or statistics: layperson testing, authority testimony, individual responsibility, and a social contract with complexity. By utilizing this framework to analyze and understand advances in technology, particularly within the latest black boxes created by certain machine learning techniques, individuals can remain epistemically honest and responsible users of ever-advancing technological advancements."

---

3. Interactive Story Generation

Faeze Brahman, Snigdha Chaturvedi

Automatic story generation is a challenging problem which requires automatically generating causally related and logical sequences of events about a topic. Previous works in this domain generate entire story at once with limited input from the user. We instead focus on the task of interactive story generation, where the user provides the model a mid-level sentence abstraction in the form of cue phrases. This provides an interactive interface for humans to supervise the story generation process. We propose two novel content-introducing approaches to incorporate this additional information. Experimental results based on automatic as well as human evaluations show that the proposed models improve over baselines in generating more coherent, on-topic and personalized stories."

---

4. IGM-Vis: Analyzing Intergalactic and Circumgalactic Medium Absorption Using Quasar Sightlines in a Cosmic Web Context

Joseph N. Burchett, David Abramov, Jasmine Otto, Cassia Artanegara, J. Xavier Prochaska, Angus G. Forbes

We introduce IGM-Vis, a novel astrophysics visualization and data analysis application for investigating galaxies and the gas that surrounds them in context with their larger scale environment, the Cosmic Web. Environment is an important factor in the evolution of galaxies from actively forming stars to quiescent states with little, if any, discernible star formation activity. The gaseous halos of galaxies (the circumgalactic medium, or CGM) play a critical role in their evolution, because the gas necessary to fuel star formation and any gas expelled from widely observed galactic winds must encounter this interface region between galaxies and the intergalactic medium (IGM). We present a taxonomy of tasks typically employed in IGM/CGM studies informed by a survey of astrophysicists at various career levels, and demonstrate how these tasks are facilitated via the use of our visualization software. Finally, we evaluate the effectiveness of IGM-Vis through two in-depth use cases that depict real-world analysis sessions that use IGM/CGM data.

---

5. BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2

Melissa Cline, Gunnar Ratsch, Amanda Spurdle, Amy Coffin, Rob Currie, Benedict Paten, BRCA Challenge Consortium

Pathogenic variation in BRCA1 and BRCA2 can increase a woman's lifetime risk of breast cancer from the population average of 12% to 65% or higher. Additionally, heritable breast cancers are more aggressive, and strike at earlier ages. Genetic testing is now allowing more women to understand and manage their heritable cancer risk, but the effectiveness of genetic testing is limited by our large gaps in genetic knowledge: even in the well-characterized BRCA genes, upwards of 40% of all variants are of uncertain clinical significance. One reason for this problem is the difficulty in sharing genetic data. To address this issue, the Global Alliance for Genomics and Health (GA4GH) launched the BRCA Challenge to develop efficient and effective public data aggregation on two high penetrance genes, and to lay the technical, legal, and cultural foundations necessary for widespread data sharing. BRCA Exchange, the first work product of the BRCA Challenge, is the largest public source of BRCA variation data, and supports variant interpretation through data aggregation, in silico prediction, and text mining. We are adding new features regularly to BRCA Exchange, and plan to soon expand the set of genes.

---

6. Interactive Canvas Style Transfer

Mahika Dubey, Jasmine Otto

This paper introduces in-browser applications for the application of style-transfer brushes onto an image. We present two distinct approaches to the creation of an application that invites ’casual creators,’ and other nontechnical users to interact with pre-trained deep convolutional neural networks to co-create customized art. In the first approach, called Magic Markers, we give the users an experience that mimics painting with a brush on a canvas, such that they are able to ’paint’ a style onto parts of their image through intuitive mouse selection and dragging over a canvas object. The second approach, Compositing Stamps, uses a real-time transfer method for applying style ’filters’ to selected rectangular portions of an image. This process reveals to users some interesting features of the style transfer functions such as border artifacts, spatial stability, and multi-layering of different styles. The two applications provide new perspectives on a well-known algorithmic process, and enhances intuition for its expressive range, or lacks therein by using traditional paintings and novel data visualization art pieces for generating styles.

7. Aligning Product Categories using Anchor Products

Varun R Embar, Golnoosh Farnadi, Jay Pujara, Lise Getoor

E-commerce sites group similar products into categories, and these categories are further organized in a taxonomy. Since different sites have different products and cater to a variety of shoppers, the taxonomies differ both in the categorization of products and the textual representation used for these categories. In this paper, we propose a technique to align categories across sites, which is useful information to have in product graphs. We use breadcrumbs present on the product pages to infer a site's taxonomy. We generate a list of candidate category pairs for alignment using anchor products - products present in two or more sites. We use multiple similarity and distance metrics to compare these candidates. To generate the final set of alignments, we propose a model that combines these metrics with a set of structural constraints. The model is based on probabilistic soft logic (PSL), a scalable probabilistic programming framework. We run experiments on data extracted from Amazon, Ebay, Staples and Target, and show that the distance metric based on products, and the use of PSL to combine various metrics and structural constraints lead to improved alignments.

8. Deep-Packet Analytics of SCADA Network Traffic Data

Mustafa Faisal, Xi Qin, Kelvin Mai, Alvaro A. Cardenas

The security of Industrial Control Systems (ICS) has a prominent place in cyber security because ICS are commonly used in critical infrastructures, such as power grid, water and chemical plants, etc. To identify abnormal events and attacks, we are currently working on developing deep-packet inspection tools for industrial networks to build models of the behavior of these systems based on network traffic data. Our models have been applied to network communication datasets from a testbed of equipment used in real-world power generation substations. Our experiments are focusing on two well-known industrial network protocols: (1) Modbus/TCP and (2) DNP3. Our anomaly detection system is able to present the informative discrete state diagrams of ICS and spot abnormal states or transitions through unseen states or transition probabilities. With these functions, the goal of our system is to alert the operators of new unusual behavior in their networks and provide them with enough contextual information so that they can remediate these issues.

9. Art I Don't Like

Sarah Frost, Manu Mathew Thomas, Angus Forbes

Art I Don’t Like is a Web-based interactive art experience that provides personalized content to users and emphasizes the introduction of disparate content. We suggest a new “anti-recommender” system that provides content that is aesthetically related in terms of low-level features but challenges the implied conceptual frameworks indicated by initial user selections. Furthermore, we demonstrate an application of recommender technologies to visual art in an effort to expose users to a broad range of art genres. We present details of a prototype implementation trained on a subset of the WikiArt dataset, consisting of 52,000 images of art from 14th- to 20th-century European painters, along with feedback from users. Art I Don’t Like is on the web at http://www.artidontlike.com.

10. Learning on Label Scarce Data

Akul Goyal, Yang Liu

Majority of data available is often unlabeled or very small part of it is labelled. Training accurate classifier becomes challenging with such variability found within the data. We propose a unique method which uses a two step process to achieving significant accuracy on data with little to no labels. Our method uses an ensemble of classifiers and aggregates their results through variational inference belief propagation. We then take this aggregation and run it through a modified adaboosting algorithm which corrects for the noise within the data. Our results are able to come close to the accuracy of a classifier running on a fully labeled dataset.

11. The power of pivoting in counting cliques

Shweta Jain, C. Seshadhri

Cliques counts are important in the analysis of large graphs. While there are several methods to count cliques up to size 12, these do not scale for larger cliques, despite using parallelization. Even approximate methods end up needing a huge amount of space and/or time.

We present a non-parallel, exact clique counting algorithm called BKP that takes a *fraction* of the time taken by other state of the art algorithms and gives counts of *all* cliques. The algorithm stems from a shockingly simple but elegant observation about the pivoted version of the famous Bron-Kerbosch algorithm for counting maximal cliques. For counting cliques in a graph with few million vertices and edges, state of the art methods for counting k-cliques for k=13 had not terminated even after 8 hours, where BKP counted all cliques in 30 seconds."

---

12. A self-exciting point process based on non-parametric Bayesian approach: modeling the pattern of occurrences of earthquakes

Hyotae Kim, Athanasios Kottas

"In seismology, one of interest is the pattern of occurrences of earthquakes, which can help the seismologist understand and predict the pattern of earthquakes occurred by time. To obtain such patterned information, a stochastic process called the Hawkes process, is used in this paper. The Hawkes process modeling yields a big advantage that makes it possible to analyze separately the main shock and aftershock. They are provided by the intensity function characterizing the processes.

We introduce a non-parametric Bayesian approach – mixture of Erlang distributions -- for inference about the intensity function of the process. The proposed approach is a mixture of Erlang distributions which have a scale parameter in common where the shape parameters of the distributions are determined – j-th mixture component of distribution has the shape parameter, j. From the perspective of parameter estimation, our proposed mixture model has a huge benefit, that is, we need to estimate only the common scale parameter and weights while, in general, J components mixture model has at least J parameters and its weights.

When we define the weight of each component, we use a function assigned a non-parametric process prior that gives the mixture model much flexibility. Computational tractability is another benefit of our mixture model. We will elaborate with earthquake data set (available online) the mixture model and the stochastic process that our model is based on."

---

13. Multi-Scale Shotgun Stochastic Search for Large Spatial Datasets

Daniel Kirsner and Bruno Sansó

Large spatial datasets often have small scale features that only occur in part of the space, coupled with large scale features across the entire space. We develop a multi-scale spatial kernel convolution model where small scale local features are captured by high resolution knots while lower resolution terms are used to describe large scale features. To achieve parsimony and explicitly identify the subdomains of the space that exhibit fine scale attributes, we develop a form of shotgun stochastic search coupled with a novel stochastic process prior that results in spatially varying resolution. This spatially varying resolution allows the statistician to produce graphics that highlight the regions with small scale local features.

14. Classification Cube

Avital Meshi

“Classification Cube” is an interactive art installation which invites viewers to become engaged with a system of Machine Learning (ML) classifier algorithms. The immersive space allows viewers to see for themselves how their bodies are being viewed and classified. It also shows a pre-recorded classification video of a diverse group of animated human avatars. The classification process, derived from open source algorithms, recognizes faces and classify bodies by their gender, age, emotion and kinetics. While inside the space, viewers quickly learn that the system is limited to identify their bodies correctly. They also see that mis-identification is true to other bodies as well and not just to their own specific body. This experience exposes both the social oppression and the opportunities that come with ML classifications of the human body. On one hand, based on a fixed data set to be trained upon and a limited array of labels to choose from, it becomes clear that the system is not diverse enough to include all bodies, identities and behaviors. On the other hand, viewers understand that they can manipulate the system by performing to it. Spending more time and becoming familiar with these algorithms, brings the awareness that one has the power to change the way the system views and classifies a body. This notion brings about the opportunity to control the way by which machine learning classifiers see us.

15. Efficiently Counting All 5-vertex Subgraph Orbits for Vertices

Noujan Pashanasangi, C. Seshadhri, Akul Goyal

Node orbit counting is a fundamental problem in network analysis particularly in bioinformatics and social networks. There are only a few approaches known for orbit counting, but none of them scale to graphs with tens of millions of edges.

We present an algorithmic framework that for each node in a graph and each subgraphs pattern up to size five, it counts the number of times the node is incident to the subgraph, and how many of those times it plays a specific role in the subgraph. These counts are called node orbit counts, and they represent useful information about local structure of the graph. Our algorithm also computes for each edge in the graph, the number of times it plays the role of a specific edge in subgraph patterns of size 4 or less. Our framework is based on cutting a pattern into smaller ones, and using counts for orbits in smaller patterns to get count of orbits in larger patterns.

We empirically evaluate our algorithm on real-world graphs and show that it can compute node orbit counts for graphs with millions of edges in minutes.

---

16. Entity Resolution on Online Games using Collective Social Behavior

Connor Pryor, Vibin Vijay, Eriq Augustine, & Lise Getoor

Entity resolution (ER) is the problem of extracting, matching, and resolving references, into their underlying entities. This process is a crucial and expensive step in many domains such as data mining, database management, information retrieval, machine learning, natural language processing, statistics, and many more. Classical techniques for Entity Resolution generally make an independence assumption that resolves entities using some pairwise similarity function. These functions tend to use the reference's local attributes to compute some kind of score for resolution. This independence assumption severs all information between references which loses valuable information about the world. Structured connections such as relationships, friendships, and proximity can provide valuable information to resolve tricky entities. This paper will explore such a task and improve further by incorporating a collective social behavioral representation for each reference. As a proof of concept entity resolution will be performed over two chess websites in which profiles will be resolved to the underlying player entity.

---

17. Understanding the Factors that Affect Wellbeing after Intervention

Angela Ramirez, Steve Whittaker

Individual's language usage has been shown to provide insight into their mental state. This began with Pennebaker's work who created a way to computationally analyze journal entries using Linguistic Inquiry and Word Count (LIWC). We used LIWC categories to analyze the text from three different well being applications EmotiCal, Echo, and MoodAdapter. Using multiple linear regressions, we were able to understand whether topics (family, friend, religion, etc.), function words, emotions, time, traits, or original disposition influenced a change in well being. We also created a binary classifier that would predict whether a user ended the intervention with a high or low dispositional state using the same criteria. By doing so, we found that the biggest indicator wasn’t a linguistic category, but often it was the user's prior dispositional state.

---

18. Can Neural Generators for Dialogue Learn Sentence Planning and Discourse Structuring?

Lena Reed, Shereen Oraby, Marilyn Walker

Responses in task-oriented dialogue systems often realize multiple propositions whose ultimate form depends on the use of sentence planning and discourse structuring operations. For example a recommendation may consist of an explicitly evaluative utterance e.g. Chanpen Thai is the best option, along with content related by the justification discourse relation, e.g. It has great food and service, that combines multiple propositions into a single phrase. While neural generation methods integrate sentence planning and surface realization in one end-to-end learning framework, previous work has not shown that neural generators can: (1) perform common sentence planning and discourse structuring operations; (2) make decisions as to whether to realize content in a single sentence or over multiple sentences; (3) generalize sentence planning and discourse relation operations beyond what was seen in training. We systematically create large training corpora that exhibit particular sentence planning operations and then test neural models to see what they learn. We compare models without explicit latent variables for sentence planning with ones that provide explicit supervision during training. We show that only the models with additional supervision can reproduce sentence planning and discourse operations and generalize to situations unseen in training.

---

19. Comparing weight learning techniques of Probabilistic Template Languages

Shresta Bellary Seetharam, Vihang Godbole, Eriq Augustine, & Lise Getoor

This poster presents a summary of weight learning experiments performed on Probabilistic Templating Languages, PSL, and Tuffy.

---

20. Global Edge Data Management

Abhishek Singh, Natasha Mittal, Holly Casaletto, Faisal Nawab

We present three ideas for data management on Edge Networks: An Edge Database, Cooperative LSM and Minority Consensus.

21. Bayesian Mixed Effect Sparse Tensor Response Regression Model with Joint Estimation of Activation and Connectivity

Dan Spencer, Rajarshi Guhaniyogi, Raquel Prado

Brain activation and connectivity analyses in task-based functional magnetic resonance imaging (fMRI) experiments with multiple subjects are currently at the forefront of data-driven neuroscience. In such experiments, interest often lies in understanding activation of brain voxels due to external stimuli and strong association or connectivity between the measurements on a set of pre-specified group of brain voxels, also known as regions of interest (ROI). This article proposes a joint Bayesian additive mixed modeling framework that simultaneously assesses brain activation and connectivity patterns from multiple subjects. In particular, fMRI measurements from each individual obtained in the form of a multi-dimensional array/tensor at each time are regressed on functions of the stimuli. We impose a low-rank PARAFAC decomposition on the tensor regression coefficients corresponding to the stimuli to achieve parsimony. Multiway stick breaking shrinkage priors are employed to infer activation patterns and associated uncertainties in each voxel. Further, the model introduces region specific random effects which are jointly modeled with a Bayesian Gaussian graphical prior to account for the connectivity among pairs of ROIs. Empirical investigations under various simulation studies demonstrate the effectiveness of the method as a tool to simultaneously assess brain activation and connectivity. The method is then applied to a multi-subject fMRI dataset from a balloon-analog risk-taking experiment in order to make inference about how the brain processes risk.

22. Lifted Hinge-Loss Markov Random Fields

Sriram Srinivasan, Behrouz Babaki, Golnoosh Farnadi, Lise Getoor

Statistical relational learning models are powerful tools that combine ideas from first-order logic with probabilistic graphical models to represent complex dependencies. Despite their success in encoding large problems with a compact set of weighted rules, performing inference over these models is often challenging. In this paper, we show how to effectively combine two powerful ideas for scaling inference for large graphical models. The first idea, lifted inference, is a well studied approach to speeding up inference in graphical models by exploiting symmetries in the underlying problem. The second idea is to frame Maximum a posteriori (MAP) inference as a convex optimization problem and use alternating direction method of multipliers (ADMM) to solve the problem in parallel. A well-studied relaxation to the combinatorial optimization problem defined for logical Markov random fields gives rise to a hinge-loss Markov random field (HLMRF) for which MAP inference is a convex optimization problem. We show how the formalism introduced for coloring weighted bipartite graphs using a color refinement algorithm can be integrated with the ADMM optimization technique to take advantage of the sparse dependency structures of HLMRFs. Our proposed approach, lifted hinge-loss Markov random fields (LHL-MRFs), preserves the structure of the original problem after lifting and solves lifted inference as distributed convex optimization with ADMM. In our empirical evaluation on real-world problems, we observe up to a three times speed up in inference over HL-MRFs.

---

23. Using Expression-based Variant Impact Phenotyping to Characterize Cancer Variants

Thornton AM, Giannakis M, Kim E ,Boehm JS, Hahn W, Garraway L, Berger AH, Brooks AN

While advancements in genome sequencing have identified millions of somatic mutations in cancer, their functional impact is poorly understood. Recently, we have presented the expression-based variant impact phenotyping (eVIP) method to use gene expression data to characterize the function of mutations. The eVIP method uses a decision tree-based algorithm to predict the functional impact of somatic variants by comparing gene expression signatures induced by introduction of wild-type versus mutant cDNAs in cell lines. The method distinguishes between variants that are gain-of-function, loss-of-function, or neutral by comparing the gene expression consistency among replicate introductions of the variants and the wild-type alleles. We developed eVIP using the L1000 gene expression assay; however, the assay only surveys the expression of one-thousand landmark transcripts. Here, we present the application of eVIP on RNA-sequencing data and for pathway analysis. The study using RNA-sequencing examined the impact of two frameshift mutations in RNF43, which is commonly mutated in colorectal and endometrial cancers. We wanted to determine if these two hotspot mutations were functioning similarly. The eVIP algorithm predicted the RNF43 G659fs variant to have a change of function and the RNF43 R117fs variant to cause a loss of function, which is consistent with the R117fs mutation occurring earlier in the gene. In the RNF43 G659fs variant, eVIP pathway analysis identified specific pathways that were gain-of-function, such as the interferon gamma response. The eVIP method is an important step in overcoming the current challenge of variant interpretation in the implementation of precision medicine.

---

24. MOBA Item Recommendation

Jason Ting, Eriq Augustine, & Lise Getoor

Multiplayer Online Battle Arena (MOBA) games have been very popular for the past 10+ years from DOTA 2 to League of Legends. Each match consist of two teams of five players each controlling a different hero. The purpose of this game is to take down the enemy's base. Throughout the game, players will purchase different items using in game currency depending on ally heroes, enemy heroes, and your own hero. Different items will power up heroes in different way. Thus, choosing a good combination of items is essential for the team’s success. In this project, we propose a recommendation system using Probabilistic Soft Logic (PSL) for selecting the optimal items.

---

25. Petfinder: Predict a Pet's Adoption Speed

Nikhil Varghese, Andy Vitek, Zekun Zhao

The goal of this project is to improve animal adoption rates and to guide shelters and rescuers to improve their pet profiles’ appeal, reducing animal suffering. In order to do
this, we constructed a machine learning model which predicts how fast a pet will be adopted, based on its characteristics, which is a supervised classification problem.
This challenge is very interesting to us because it is not a simple image classification task, and there are other features which could potentially be even more important
than how the pet looks like.

---

26. Signal Processing of EEG data for Deep Neural Networks

Christopher Villalpando, Michael Briden

We propose a novel technique for pre-processing raw electroencephalogram signals for emerging machine learning algorithms that excel at pattern recognition. Our technique condenses large amounts of information into a relatively small 3-dimensional structure consisting of spatial, temporal and frequency domains. In this poster we show our previous results as well as the current progress.

---

27. Improving Tsunami Early Warning with Deep Learning

Dimitri Voytan, Thorne Lay, Emily Brodsky

Large earthquakes that rupture at the interface between a subducting oceanic and continental plate are potentially tsunamigenic. These earthquakes vary substantially in rupture depth, with some rupturing at depths of several tens of kilometers beneath the elevation of the sea floor and others rupturing to the oceanic trench. The latter case, earthquakes with shallow slip, can substantially enhance the size of a resulting tsunami relative to a deeper earthquake of the same magnitude, posing a significant hazard to coastal communities. For subduction zone earthquakes with either shallow or deep slip, the upgoing P-wave generated by the rupture process refracts at the boundary between the oceanic trench and seafloor converting into a reverberating water multiple called pWP. When slip during the earthquake extends shallow, this water reverberation is enhanced relative to earthquakes with deeper slip, and results in 'ringing' in the seismogram well after the expected duration of energy arriving as a result of the rupture process. When filtered in the 20 second to 5 second passpand, this ringing is easily distinguishable even to non specialists. This readily apparent characteristic of seismograms from earthquakes with shallow slip motivates the use of deep learning to distinguish between these two classes of earthquake. We train a convolutional neural network on seismograms from 40 major (Mw > 7) megathrust earthquakes to detect whether an earthquake has shallow or deep slip. Initial cross-validation results show that the neural network performs similarly to a currently proposed parametric method for event discrimination, with the advantage or autonomy.

---

28. CruzAffect: a feature-rich approach to characterize happiness

Jiaqi Wu, Ryan Compton, Geetanjali Rakshit, Marilyn Walker, Pranav Anand, and Steve Whittaker

We present our system, CruzAffect, for the CL-Aff Shared Task 2019. CruzAffect consists of several types of robust and efficient models for affective classification tasks. We utilize both traditional classifiers, such as XGBoosted Forest, as well as a deep learning Convolutional Neural Networks (CNN) classifier. We explore rich feature sets such as syntactic features, emotional features, and profile features, and utilize several sentiment lexicons, to discover essential indicators of social involvement and control that a subject might exercise in their happy moments, as described in textual snippets from the HappyDB database. The data comes with a labeled set (10K), and a larger unlabeled set (70K). We therefore use supervised methods on the 10K dataset, and a bootstrapped semi-supervised approach for the 70K. We evaluate these models for binary classification of agency and social labels (Task 1), as well as multi-class prediction for concepts labels (Task 2). We obtain promising results on the held-out data, suggesting that the proposed feature sets effectively represent the data for affective classification tasks. We also build concepts models that discover general themes recurring in happy moments. Our results indicate that generic characteristics are shared between the classes of agency, social and concepts, suggesting it should be possible to build general models for affective classification tasks.

29. NeuroCave

Ran Xu, Manu Mathew Thomas, Oskar Elek, Olusola Alade Ajilore, Angus Forbes

The human brain consists of millions of neural connections called connectomes. Neuroscientists and researchers find it difficult to interpret these connections due to the high volume. NeuroCave is an immersive interactive 3D visualization tool build for neuroscientists to inspect structural and functional time series connectome datasets. With this tool, a neuroscientist can interact with the connectome—either in a standard desktop environment or in a VR space. Our application allows researchers to upload connectome data and explore connections in different representations and regions of the brain. We introduce a new design for an overlay comparison to compare connections in two different time series connectomes. Visual clutter is mitigated using a state-of-the-art edge bundling technique and through an interactive layout strategy.

30. The Game of Protocols: Using Reinforcement Learning to Improve the Performance of Channel Access Schemes

Molly Zhang, J. J. Garcia-Luna-Aceves, Luca de Alfaro

We are using machine learning (ML) to allow medium-access control (MAC) protocols achieve higher networking bandwidth, lower delays, and better fairness. MAC protocols regulate how nodes access the transmission medium in a network. Each node follows rules to transmit data in order to avoid collision and to maximize the speed at which the data packets are transmitted. Historically, there have been both schedule-based and contention-based protocols, such as TDMA and ALOHA. A node in TDMA transmits at a regular interval regardless of what happens in a network, and a node using a contention-based MAC protocol attempts to transmit with some probability that is independent of the channel utilization. We are studying new families of MAC protocols that use ML to achieve backwards compatibility with such existing protocols as TDMA and ALOHA, and also attain much higher efficiency and fair sharing of the bandwidth.