My Research

Machine Learning for complex social, technological, and biological systems
New Paradigm For Discovery

Massive data: Observe “invisible” patterns
My Research: End to End

New computer science:
New algorithms, new ML, new DM.

New science:
Insights into humanity, new public policy recommendations.

Jure Leskovec, Stanford University
My Group’s Research

- Computer Systems
- Theory and Algorithms
- Applications & Industry
- Social Sciences
- Medicine
- Public policy
- Web
- SNAP group

Jure Leskovec, Stanford University
Fruitful Collaborations
New Ways of Thinking

Working on real-world problems leads to new ways of thinking:

- Pure algorithmic questions turn out not to be so important
- More important is methodology and understanding human behaviors
Our Group: Three Themes

1) ML for improving human health
   - How smartphones allow us to study obesity pandemic

2) ML for human decision making
   - Why are criminal court judges making mistakes?

3) Deep learning for complex networks
   - What are the limits of predictability in complex networks
Large-scale sensing to improve human health
State of the Art in Health Research

- Health studies & data:
  - High financial cost
  - Limited scale and resolution
    - Few people agree to research study
  - Short-term
    - Few days of observation
  - Biases from self-reporting
  - Inability to blind the user/patient
Big Opportunity

- Growing popularity of wearables
  - Smartphones & wearable sensors
  - Activity and health data across millions of individuals

- How can wearables be used to gain insights into human health?
How much do we exercise?

Exercise is extremely important for health [Lee et al., 2012]

But we do not know how much activity people get!

- WHO: 5-54% of Germans get insufficient exercise. No data for Switzerland and Israel.
Data: Azumio & UnderArmour

- 5.6 million users from over 120 countries
- 791 million actions recorded
- 160M days of steps tracking
  - >230B data points
  - Includes online social network

- Demographics fairly representative (USA)
  - Gender: 50.2% female (publ.: 50.8%)
  - Age: Median 35 years (publ.: 37 years)
  - Obesity: 28.8% (published: 29-35%)
In-memory Analytics

- System for large scale in-memory analytics
  - Fast execution times: Interactive use
  - Support for several data representations
    - Transformations between tables and graphs
  - Large number of ready-to-use algorithms

Jure Leskovec, Stanford University
Worldwide Activity

For most of these countries WHO provides no official estimate of physical activity
Inequality of Physical Activity

How is activity distributed among the population?

Gini index of the activity distribution: How (un)evenly is activity distributed?

Difference in means

Difference in variance

Jure Leskovec, Stanford University
Obesity vs. Activity Inequality

\[ R^2 = 0.65 \] (vs. 0.4 for avg. activity)
Gender activity gap contributes to inequality.
Questions and Solutions

How could we combat activity inequality?

- **Built environment**
  - Change how we design cities

- **Social networks**
  - Use social networks to create effective support communities
Environment & Walkability

Walkability: Measure of how friendly an area is to walking (based on physical environment)

- New York, NY
- San Antonio, TX
Benefits from Walkability

![Graph showing the benefits of walkability for different age groups and BMI categories](image-url)
Can Social Networks Help?

Alice → friend → Bob
Impact of Social Networks

Fundamental challenge: We never observe what would happen if the friendship edge would not be created.
Causal Identification

New method for causal identification of social influence: A natural experiment exploiting delay of friendship acceptance

Direct edge accept:  
Delayed accept:

The difference isolates social influence!

Jure Leskovec, Stanford University
Social Networks Influence Activity

Online social networks influence physical activity!

- Effect is strong (~350 steps per day)
- Effect is long-term (~5 months)
Next Steps

ML for improving human lives:

- Mental-health counseling
  - Millions of transcripts of counselor-patient discussions and outcomes
- Food intake
  - Identifying food deserts
  - Helping people eat healthier
Machine Learning and Human Decisions
Decision Problems

Medicine
What treatment do we recommend to a patient?

Education
Is a student at risk of dropping out?

Criminal justice
Will a person commit a crime if let out on bail?

Today humans make decisions.
How can machines help?
Human vs. Machines

- Information asymmetry/advantage
  - Human judges see many things that are hard to quantify

- Behavioral biases in decision making and inference
  - Human judges may induce subconscious or intentional biases
Criminal Court Bail Decisions

- U.S. police makes >12M arrests/year
  Where do people wait for trial?
- Judge decides whether to release or not
- Defendant out on bail can behave badly:
  - Fail to appear at the trial
  - Commit another crime
- Judge’s question: Will defendant misbehave while out on bail?

Judge is making a prediction!
Typical ML Setup

Training:
- Based on released defendants build an ML model

Testing:
- Assume real judge releases $k$ defendants
- For every defendant predict $P\text{(crime)}$
- Sort by increasing $P\text{(crime)}$ and “release” top $k$
- Compare crime rate of judge vs. model
What Features to Use

Only administrative variables
- Common across regions
- No “gaming the system” problem

Only variables that are legal to use
- No race, gender

- Age at first arrest
- Times sentenced residential correction
- Level of charge
- Number of active warrants
- Number of misdemeanor cases
- Number of past revocations
- Current charge domestic violence
- Is first arrest
- Prior jail sentence
- Prior prison sentence
- Employed at first arrest
- Currently on supervision
- Prior supervision within 10 years
- Arrest for new offense while on bond
- Has active warrant
- Has active misdemeanor warrant
- Has other pending charge
- Had previous adult conviction
- Had previous adult misdemeanor conviction
- Had previous adult felony conviction
- Had previous Failure to Appear
- Had previous revocation
Problem: Selective Labels

Judge is selectively labeling the dataset

- We can only train on released people
- By jailing judge is selectively hiding labels!
Selective labels introduce bias:

- Say young people with no tattoos have no risk for crime. Judge releases them.
- But machine does not observe whether defendants have tattoos.
- So, the machine would falsely conclude that all young people do no crime.
- And falsely presume that by releasing all young people it does better than judge!
Solution: Contraction

Making lenient judges strict:

Defendants (ordered by crime probability)

What makes contraction work: Judges have similar cases (random assignment of cases)

Why is this elegant: No assumption on judge and ML using the same set of features
Predicted Failure to Appear

Judges decisions
10%

Release Rate

FTA Rate

73%
Predicted Failure to Appear

60% reduction in crime rate!

![Graph showing comparison between judges' decisions and a machine learning algorithm's predictions for Failure to Appear (FTA) rates. The graph illustrates that the machine learning algorithm reduces the FTA rate by 4.1% compared to the judges' decisions, which have a 10% FTA rate. The graph also indicates a significant reduction in crime rate, with a 73%.]

Jure Leskovec, Stanford University
Why do Judges do Poorly?

Odds are against algorithm: Judge sees many things the algorithm does not

- Offender history:
  Observable by both

- Demeanor:
  - “I looked in his eyes and saw his soul”
  - Unobservable by the algorithm

Can we diagnose judges’ mistakes and help them make better decisions?
Modeling Human Decisions

Decision of judge j on defendant i

Truth

$\mathbf{p}_{c,c}$ $\mathbf{p}_{c,n}$

$p_{n,c}$ $p_{n,n}$

Probability that j decided “no crime” to a defendant who will do “crime”

Judge j

Defendant i

True label $t_i$
Joint Confusion Model

Judge attributes

Item attributes

Judge

Defendant

Decision of judge $j$ on defendant $i$

Confusion matrix

Decision

Truth

Cluster $c_j$

Cluster $d_i$

$a_1 a_2 a_3 a_4 a_5$

$b_1 b_2 b_3 b_4 z_i$

Discover groups of judges and defendants that share similar confusion matrices

Jure Leskovec, Stanford University
Why judges make mistakes

Judges with many felony cases are stricter
- Left: Judges with many felony cases
- Right: Judges with few felony cases
Why judges make mistakes

Defendants who are single, did felonies, and moved a lot are accurately judged

Defendants who have kids are confusing to judges
Power of Algorithms

- ML diagnoses reason for bias
- Key insight:
  - More than prediction tools
  - Serve as behavioral diagnostic tools
- Next steps:
  - Human-centric ML
  - Interpretability and explainability
Deep Learning for Networks

How can we develop Deep Learning techniques for complex data?
Deep Learning for Graphs

Desiderata:

- Scale to large networks
- Generalize to new nodes
- Build a model of node embedding
  - Training phase can be slow but prediction phase should be fast
Inductive Feature Learning

Inductive node embedding

- train on one graph

Generalize to entirely unseen graphs

- generalize to an entirely new graph
The Challenge

- **Speech/text:**

- **Images:**

- **Graphs:**

  or this:
Naïve Approach

- Are we done? **Not even close.**
  - Huge number of parameters
  - No inductive learning possible
  - Poor generalization
  - Dependence on node ordering/numbering
Graph Convolution Networks

Learn how to aggregate (convolve) the information from its neighbors:

- $H^{(k+1)} = \sigma(AW^kH^{(k)})$, and $H^{(0)} = X$
- **Important**: Directly leverage input node features (e.g., attributes, degrees)
GraphSAGE

1. Sample neighborhood

2. Aggregate feature information from neighbors

3. Predict graph context and label using aggregated information

- Sample and aggregate:

\[
\begin{align*}
    h^k_{\mathcal{N}(v)} & \leftarrow \text{AGGREGATE}_k(\{h^{k-1}_u, \forall u \in \mathcal{N}(v)\}); \\
    h^k_v & \leftarrow \sigma \left( W^k \cdot \text{CONCAT} (h^{k-1}_v, h^k_{\mathcal{N}(v)}) \right)
\end{align*}
\]
GraphSAGE

- Directly leverage feature information
- Neural network that aggregates neighborhood
- Adapt to inductive setting (e.g., unsupervised loss, neighborhood sampling, minibatch optimization)
**Application: Drug Repurposing**

**Q:** Can we predict therapeutic uses of a drug?

**Insight:** A drug is likely to treat a disease if it is close to the disease in the network.

- **State of the art AUC:** <0.70
- **Our method AUC:** >0.93

Jure Leskovec, Stanford University
Application: Recommendations

Pinterest is a Giant Bipartite Graph:
- Objects belonging to collections
- Rich image and textual features
Learned Embeddings
Next Steps

- Embedding rich multimodal networks
- Answering queries beyond single link prediction
- Combining node/image features and the network structure
In Conclusion
EVERYONE STAND BACK

I KNOW COMPUTER SCIENCE
WE’RE HIRING!

Many interesting high-impact projects in Machine Learning, Social Networks, Data Bases & Systems.

Applications: Bio-medicine, Social Science
PhD Students

Claire Donnat
Mitchell Gordon
David Hallac
Emma Pierson
Geet Sethi
Himabindu Lakkaraju
Rex Ying
Tim Althoff
Will Hamilton

Post-Doctoral Fellows

David Jurgens
Marinka Zitnik
Michele Catasta
Srijan Kumar

Industry Partnerships

Collaborators

Dan Jurafsky, Linguistics, Stanford University
Christian Danescu-Miculescu-Mizil, Information Science, Cornell University
Stephen Boyd, Electrical Engineering, Stanford University
David Gleich, Computer Science, Purdue University
VS Subrahmanian, Computer Science, University of Maryland
Sarah Kunz, Medicine, Harvard University
Russ Altman, Medicine, Stanford University
Jochen Profit, Medicine, Stanford University
Eric Horvitz, Microsoft Research
Jon Kleinberg, Computer Science, Cornell University
Sendhil Mullainathan, Economics, Harvard University
Scott Delp, Bioengineering, Stanford University
Jens Ludwig, Harris Public Policy, University of Chicago

Funding

Research Staff

Stephen Bach
Peter Kacin
Rok Sosic