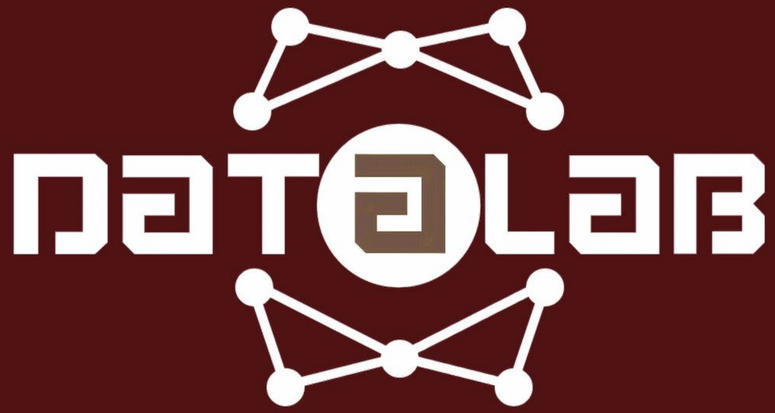


DataLab12.github.io

Data Lab @ TXST
DataLab12.github.io

Founded in 2018 by CS faculty:

Jelena Tešić, Computer Science

Objective:

- Propose new algorithms and methods to applied Data Science for Unstructured Real Data
- **Collaborate with domain experts**

TEXAS STATE UNIVERSITY

MEMBER THE TEXAS STATE UNIVERSITY SYSTEM

DataLab12.github.io

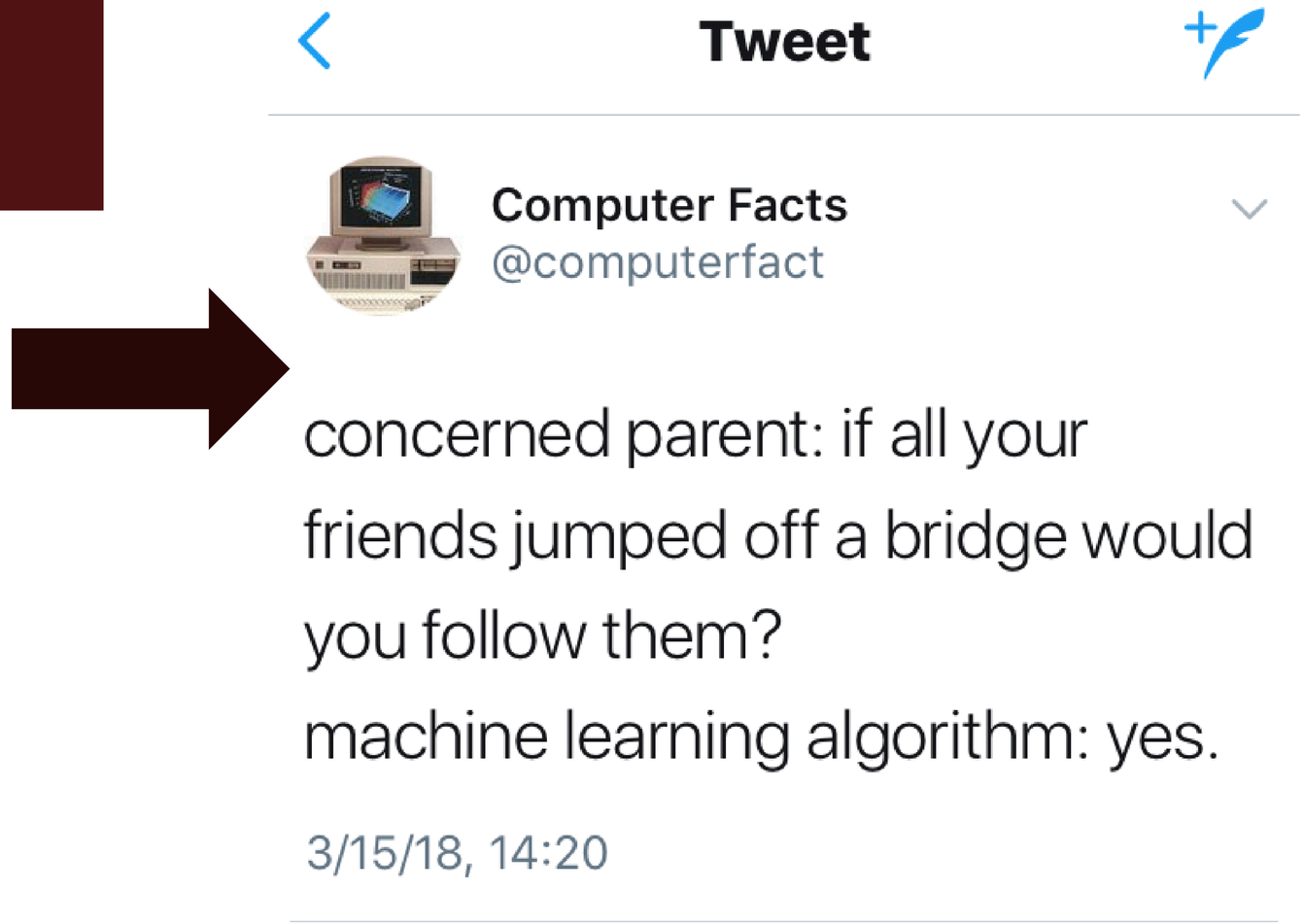
Data Lab @ TXST
DataLab12.github.io

Founded in 2018 by CS faculty:
Jelena Tešić, Computer Science

Motivation Go beyond solving incremental ML in silos

Funding

- CS Startup funding 2018 -2021
- NAVAIR funding 2018 – 2023
- THRC CHERR 2021 – 2023
- TxDot 2022- 2026
- DoE 2022-2024
- NVIDIA gifts



ABOUT ME

Managing large Multimedia Repositories, Ph.D. Thesis

- M.S. (1999) and Ph.D. (2004) degrees from Department of Electrical and Computer Engineering, University of California, Santa Barbara
- Talk by Prof. Jovanovic USC that refers to the group, about grad student experience and how it shaped what we did after: https://www.youtube.com/watch?v=9p8iJnPQSX

Research

MS generated Alt text:

- *A group of people playing baseball on a field*
- *A group of people playing a game of frisbee*

ABOUT ME: jtesic.github.io

IBM

- **IBM Watson Research (Yorktown NY) 2004 – 2009**
  - Objective image filtering in MySpace, Ontology for Multimedia, DigitalMe
  - TRECVID Challenge (Video Retrieval Systems): low shot learning, visual concept modeling

MC

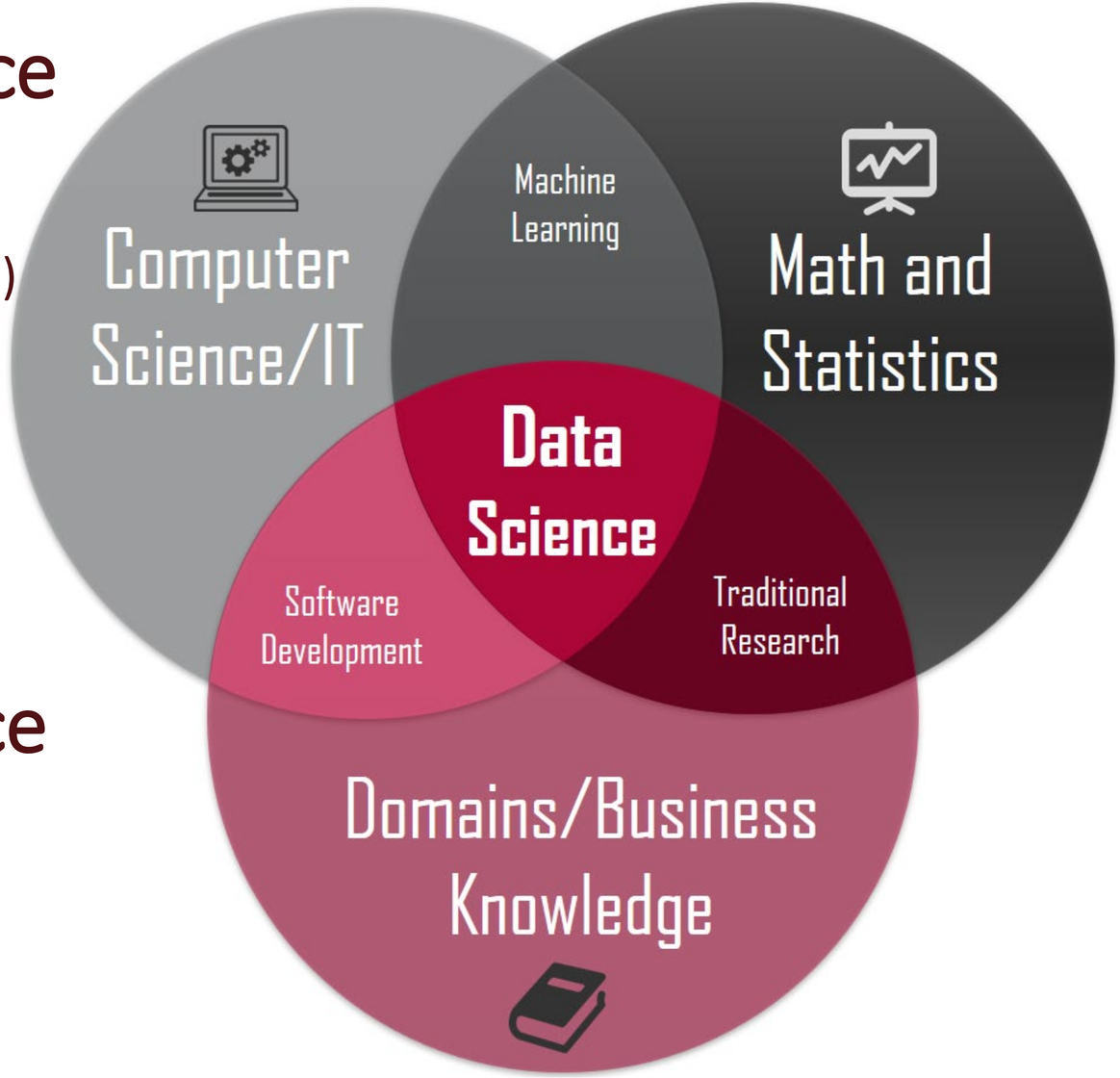
- **Mayachitra, Inc (Santa Barbara, CA) 2009 - 2018**
- Video and Image Retrieval and Analysis Tool (VIRAT) - scalable indexing, search, retrieval, data fusion
- Geo-location using image matching (FINDER)
- Activity recognition; Deep learning for Object recognition (NAVAIR) – PI since 2014

TXST

- **Computer Science Dept Data Lab @ TX State**
- https://datalab12.github.io/
- Graph network analysis at scale: graph construction from unstructured data
- Fair and unbiased data science: consensus analysis
- Applied DNN to aerial imagery, health imagery, pavement imagery, and climate modeling

Visual Data Science

- NAVAIR
- Dr. Wang, Ingram (TxDot PI)
- Dr. Faroughi, Ingram (DoE PI)



Network Data Science

- Dr. Rusnak, Math

Applied Data Science

- CHERR (Dr. Villagran)
- Dr. Feng, McCoy
- Dr. Metsis, CS
- Dr. Wang, CS

CS Courses:
Machine Learning
Data Science (Ph.D.)

DataLab12.github.io

Data lab Projects and Collaborators

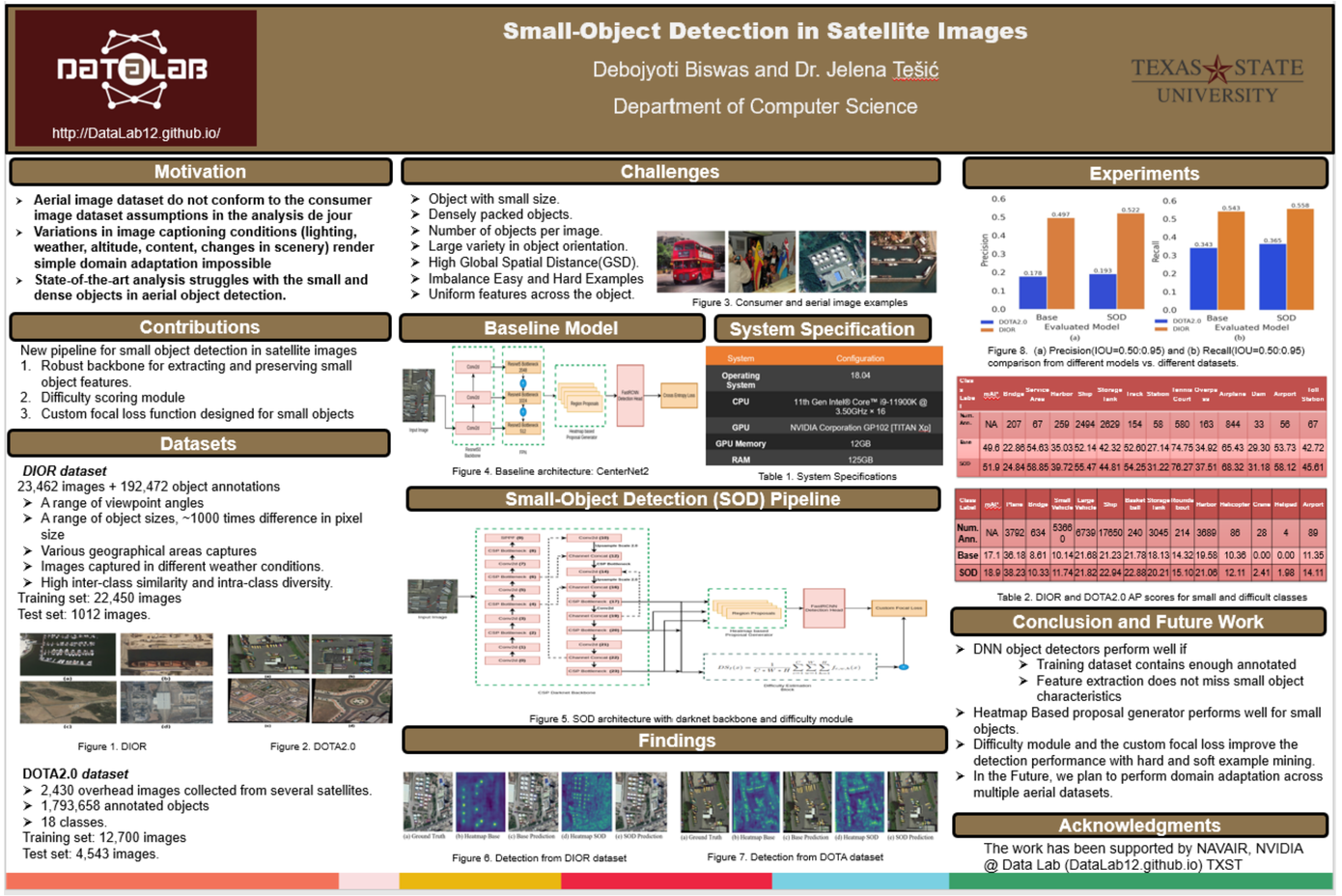
TEXAS STATE UNIVERSITY

Visual Data Science

New efficient data-driven DNN architectures

- Object localization and identification
- Activity recognition
- Segmentation
- 3D Point Cloud Modeling

Poster: Small-Object Detection in Satellite Images (Bishal, Ph.D)



A picture containing timeline
Description automatically generated

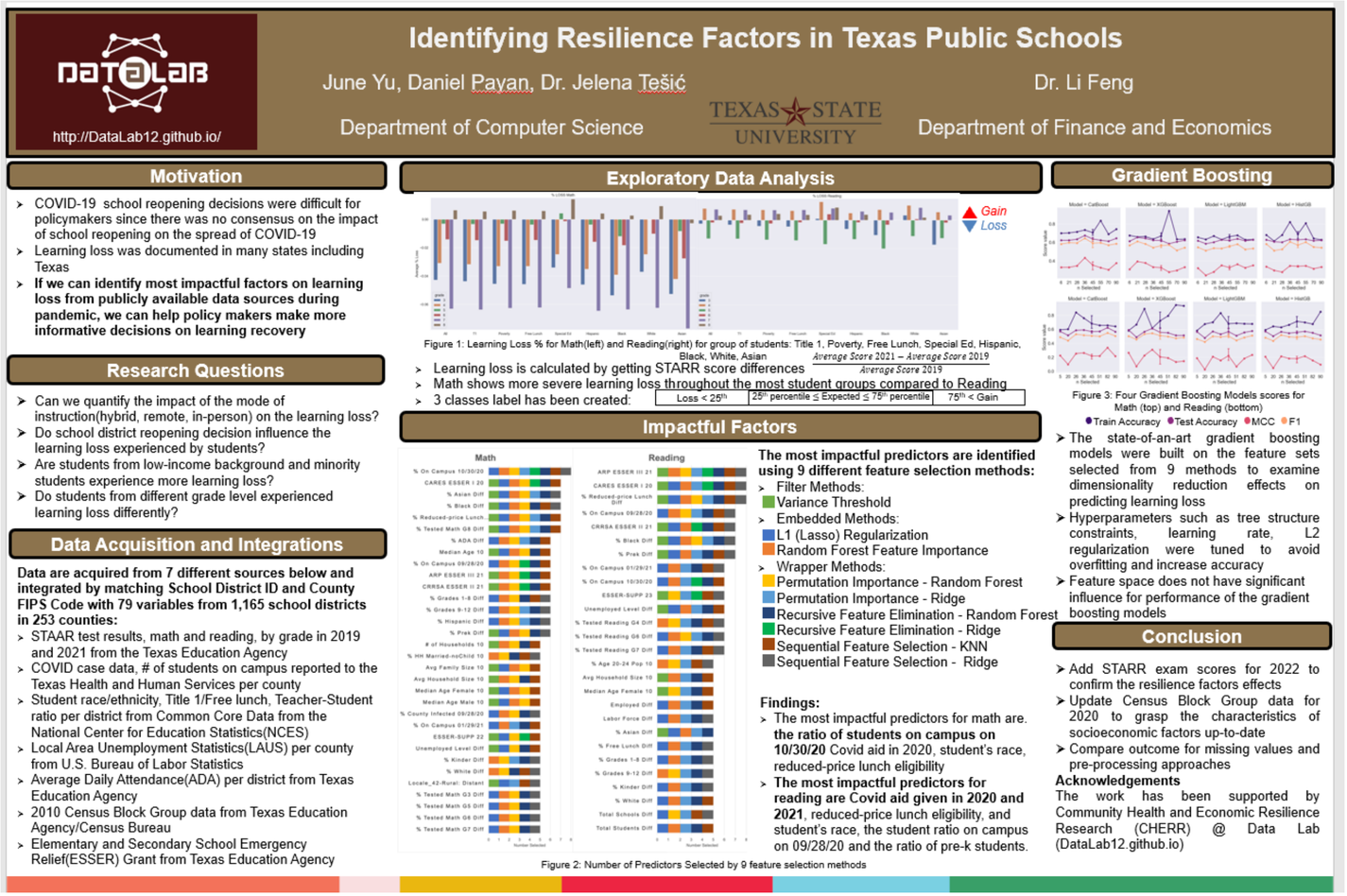
TEXAS STATE UNIVERSITY

Applied Data Science

Develop new data-driven end-to-end analytics that maximizes tabular ML advances

- **Work w domain experts to avoid GIGO**
- **Heath data, education data**

Poster: Identifying Resilience Factors in Texas Public Schools (June, M.Sc. Daniel, B.Sc.)



Identifying Resilience Factors in Texas Public Schools

http://DataLab12.github.io/

June Yu, Daniel Payan, Dr. Jelena Tešić — Department of Computer Science

Texas State University

Dr. Li Feng — Department of Finance and Economics

Motivation

- COVID-19 school reopening decisions were difficult for policymakers since there was no consensus on the impact of school reopening on the spread of COVID-19
- Learning loss was documented in many states including Texas
- **If we can identify most impactful factors on learning loss from publicly available data sources during pandemic, we can help policy makers make more informative decisions on learning recovery**

Research Questions

- Can we quantify the impact of the mode of instruction(hybrid, remote, in-person) on the learning loss?
- Do school district reopening decision influence the learning loss experienced by students?
- Are students from low-income background and minority students experience more learning loss?
- Do students from different grade level experienced learning loss differently?

Data Acquisition and Integrations

Data are acquired from 7 different sources below and integrated by matching School District ID and County FIPS Code with 79 variables from 1,165 school districts in 253 counties:

- STAAR test results, math and reading, by grade in 2019 and 2021 from the Texas Education Agency
- COVID case data, # of students on campus reported to the Texas Health and Human Services per county
- Student race/ethnicity, Title 1/Free lunch, Teacher-Student ratio per district from Common Core Data from the National Center for Education Statistics(NCES)
- Local Area Unemployment Statistics(LAUS) per county from U.S. Bureau of Labor Statistics
- Average Daily Attendance(ADA) per district from Texas Education Agency
- 2010 Census Block Group data from Texas Education Agency/Census Bureau
- Elementary and Secondary School Emergency Relief(ESSER) Grant from Texas Education Agency

Exploratory Data Analysis

Figure 1: Learning Loss % for Math(left) and Reading(right) for group of students: Title 1, Poverty, Free Lunch, Special Ed, Hispanic, Black, White, Asian

- Learning loss is calculated by getting STARR score differences $\frac{Average\ Score\ 2021 - Average\ Score\ 2019}{Average\ Score\ 2019}$
- Math shows more severe learning loss throughout the most student groups compared to Reading
- 3 classes label has been created:

Loss < 25th	25th percentile ≤ Expected ≤ 75th percentile	75th < Gain

Impactful Factors

Figure 2: Number of Predictors Selected by 9 feature selection methods

The most impactful predictors are identified using 9 different feature selection methods:

- Filter Methods:
  - Variance Threshold
- Embedded Methods:
  - L1 (Lasso) Regularization
  - Random Forest Feature Importance
- Wrapper Methods:
  - Permutation Importance - Random Forest
  - Permutation Importance - Ridge
  - Recursive Feature Elimination - Random Forest
  - Recursive Feature Elimination - Ridge
  - Sequential Feature Selection - KNN
  - Sequential Feature Selection - Ridge

Findings:

- The most impactful predictors for math are. **the ratio of students on campus on 10/30/20** Covid aid in 2020, student's race, reduced-price lunch eligibility
- **The most impactful predictors for reading are Covid aid given in 2020 and 2021**, reduced-price lunch eligibility, and student's race, the student ratio on campus on 09/28/20 and the ratio of pre-k students.

Gradient Boosting

Figure 3: Four Gradient Boosting Models scores for Math (top) and Reading (bottom)

Train Accuracy · Test Accuracy · MCC · F1

- The state-of-an-art gradient boosting models were built on the feature sets selected from 9 methods to examine dimensionality reduction effects on predicting learning loss
- Hyperparameters such as tree structure constraints, learning rate, L2 regularization were tuned to avoid overfitting and increase accuracy
- Feature space does not have significant influence for performance of the gradient boosting models

Conclusion

- Add STARR exam scores for 2022 to confirm the resilience factors effects
- Update Census Block Group data for 2020 to grasp the characteristics of socioeconomic factors up-to-date
- Compare outcome for missing values and pre-processing approaches

Acknowledgements

The work has been supported by Community Health and Economic Resilience Research (CHERR) @ Data Lab (DataLab12.github.io)

A picture containing timeline
Description automatically generated

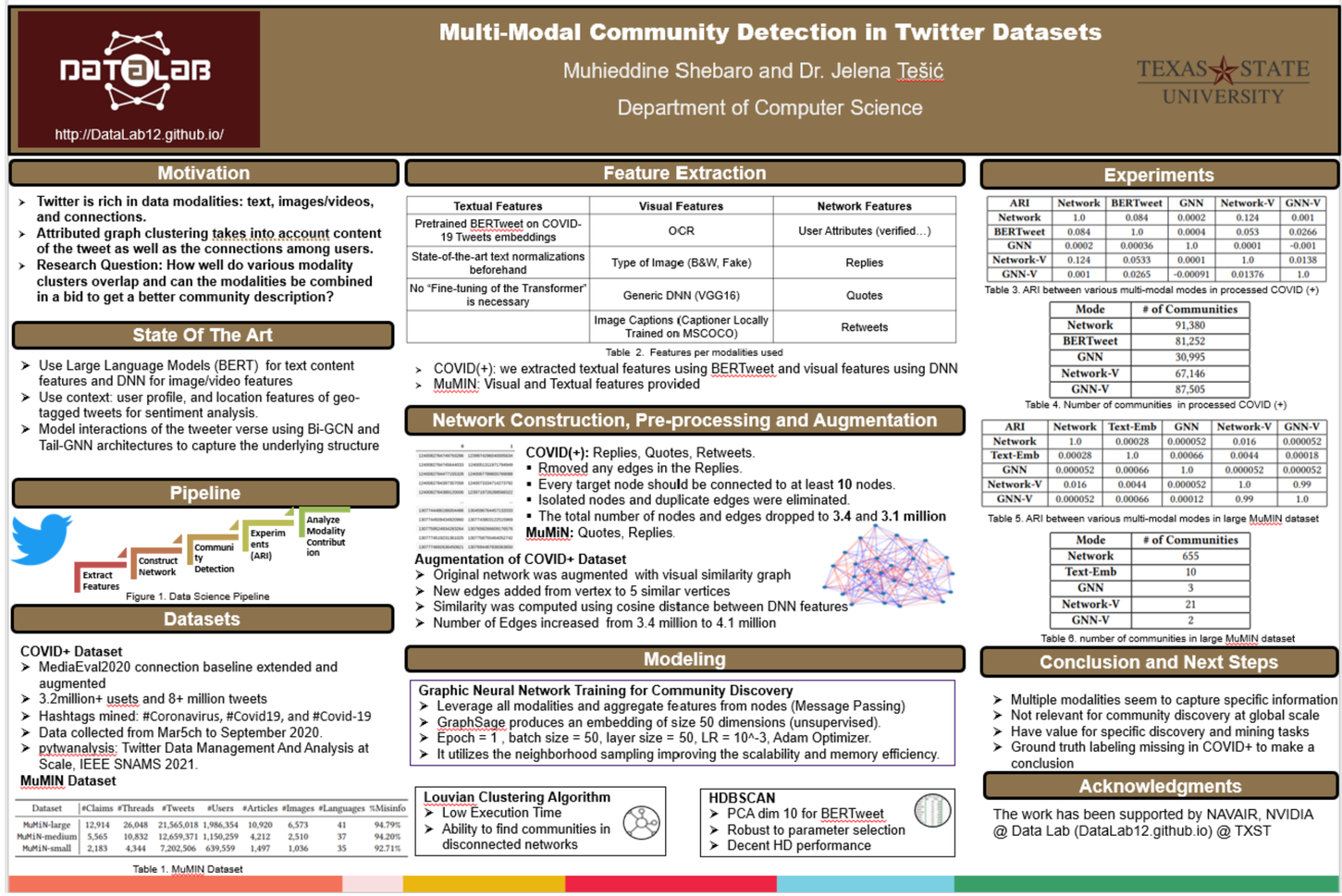
Texas State University

Network Data Science

Develop new algorithms and analytics tools for real networks

- **Where algorithms developed for synthetic data break on real networks?**
- **What is the greatest gain in network science in terms of algorithmic improvement?**

Poster: Multi-Modal Community Detection in Twitter Datasets (Mo, Ph.D.)



DataLab
http://DataLab12.github.io/

Multi-Modal Community Detection in Twitter Datasets

Muhieddine Shebaro and Dr. Jelena Tešić
Department of Computer Science
TEXAS STATE UNIVERSITY

Motivation

- Twitter is rich in data modalities: text, images/videos, and connections.
- Attributed graph clustering takes into account content of the tweet as well as the connections among users.
- Research Question: How well do various modality clusters overlap and can the modalities be combined in a bid to get a better community description?

State Of The Art

- Use Large Language Models (BERT) for text content features and DNN for image/video features
- Use context: user profile, and location features of geo-tagged tweets for sentiment analysis.
- Model interactions of the tweeter verse using Bi-GCN and Tail-GNN architectures to capture the underlying structure

Pipeline

Figure 1. Data Science Pipeline

Datasets

COVID+ Dataset

- MediaEval2020 connection baseline extended and augmented
- 3.2million+ usets and 8+ million tweets
- Hashtags mined: #Coronavirus, #Covid19, and #Covid-19
- Data collected from Mar5ch to September 2020.
- pytwanalysis: Twitter Data Management And Analysis at Scale, IEEE SNAMS 2021.

MuMIN Dataset

Dataset	#Claims	#Threads	#Tweets	#Users	#Articles	#Images	#Languages	%Misinfo
MuMiN-large	12,914	26,048	21,565,018	1,986,354	10,920	6,573	41	94.79%
MuMiN-medium	5,565	10,832	12,659,371	1,150,259	4,212	2,510	37	94.20%
MuMiN-small	2,183	4,344	7,202,506	639,559	1,497	1,036	35	92.71%

Table 1. MuMIN Dataset

Feature Extraction

Textual Features	Visual Features	Network Features
Pretrained BERTweet on COVID-19 Tweets embeddings	OCR	User Attributes (verified…)
State-of-the-art text normalizations beforehand	Type of Image (B&W, Fake)	Replies
No "Fine-tuning of the Transformer" is necessary	Generic DNN (VGG16)	Quotes
	Image Captions (Captioner Locally Trained on MSCOCO)	Retweets

Table 2. Features per modalities used

- COVID(+): we extracted textual features using BERTweet and visual features using DNN
- MuMIN: Visual and Textual features provided

Network Construction, Pre-processing and Augmentation

COVID(+): Replies, Quotes, Retweets.
- Rmoved any edges in the Replies.
- Every target node should be connected to at least **10** nodes.
- Isolated nodes and duplicate edges were eliminated.
- The total number of nodes and edges dropped to **3.4** and **3.1 million**

MuMiN: Quotes, Replies.

Augmentation of COVID+ Dataset

- Original network was augmented with visual similarity graph
- New edges added from vertex to 5 similar vertices
- Similarity was computed using cosine distance between DNN features
- Number of Edges increased from 3.4 million to 4.1 million

Modeling

Graphic Neural Network Training for Community Discovery

- Leverage all modalities and aggregate features from nodes (Message Passing)
- GraphSage produces an embedding of size 50 dimensions (unsupervised).
- Epoch = 1 , batch size = 50, layer size = 50, LR = 10^-3, Adam Optimizer.
- It utilizes the neighborhood sampling improving the scalability and memory efficiency.

Louvian Clustering Algorithm

- Low Execution Time
- Ability to find communities in disconnected networks

HDBSCAN

- PCA dim 10 for BERTweet
- Robust to parameter selection
- Decent HD performance

Experiments

ARI	Network	BERTweet	GNN	Network-V	GNN-V
Network	1.0	0.084	0.0002	0.124	0.001
BERTweet	0.084	1.0	0.0004	0.053	0.0266
GNN	0.0002	0.00036	1.0	0.0001	-0.001
Network-V	0.124	0.0533	0.0001	1.0	0.0138
GNN-V	0.001	0.0265	-0.00091	0.01376	1.0

Table 3. ARI between various multi-modal modes in processed COVID (+)

Mode	# of Communities
Network	91,380
BERTweet	81,252
GNN	30,995
Network-V	67,146
GNN-V	87,505

Table 4. Number of communities in processed COVID (+)

ARI	Network	Text-Emb	GNN	Network-V	GNN-V
Network	1.0	0.00028	0.000052	0.016	0.000052
Text-Emb	0.00028	1.0	0.00066	0.0044	0.00018
GNN	0.000052	0.00066	1.0	0.000052	0.000052
Network-V	0.016	0.0044	0.000052	1.0	0.99
GNN-V	0.000052	0.00066	0.00012	0.99	1.0

Table 5. ARI between various multi-modal modes in large MuMIN dataset

Mode	# of Communities
Network	655
Text-Emb	10
GNN	3
Network-V	21
GNN-V	2

Table 6. number of communities in large MuMIN dataset

Conclusion and Next Steps

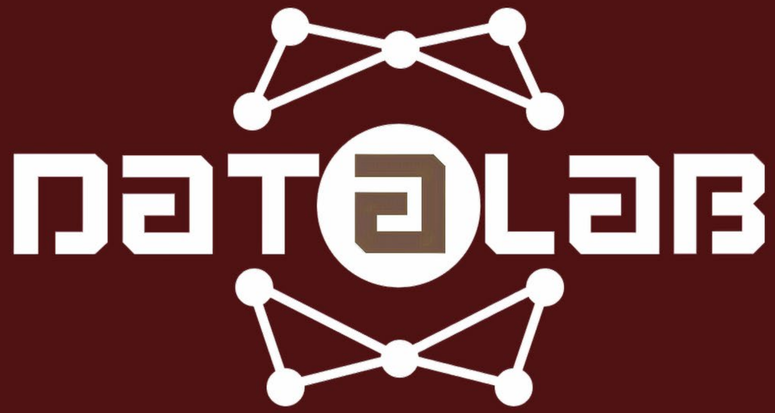
- Multiple modalities seem to capture specific information
- Not relevant for community discovery at global scale
- Have value for specific discovery and mining tasks
- Ground truth labeling missing in COVID+ to make a conclusion

Acknowledgments

The work has been supported by NAVAIR, NVIDIA @ Data Lab (DataLab12.github.io) @ TXST

A picture containing timeline
Description automatically generated

TEXAS STATE UNIVERSITY



DataLab12.github.io

Signed Graph Analysis in Real Data

Jelena Tešić, Computer Science

Lucas Rusnak, Mathematics

TEXAS STATE UNIVERSITY

MEMBER THE TEXAS STATE UNIVERSITY SYSTEM

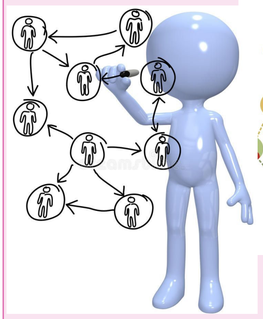
Input

Policies

Elections

Survey

Outcomes



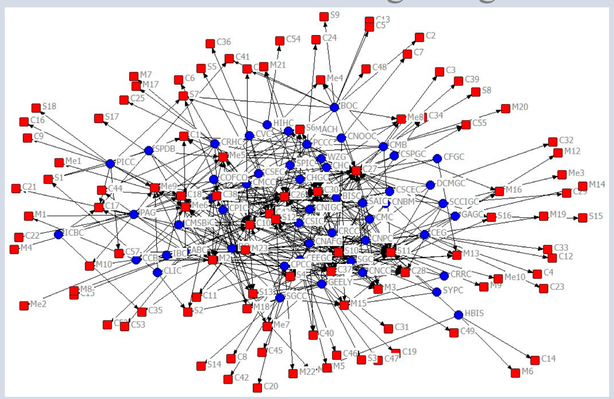
Collaboration Data

Human Resources

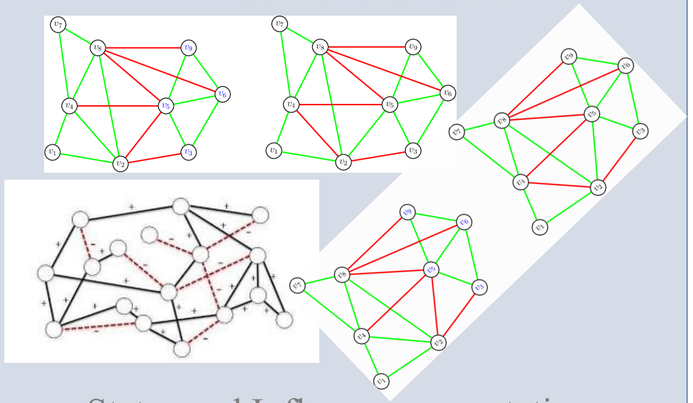
City/County/Region Data

Modeling

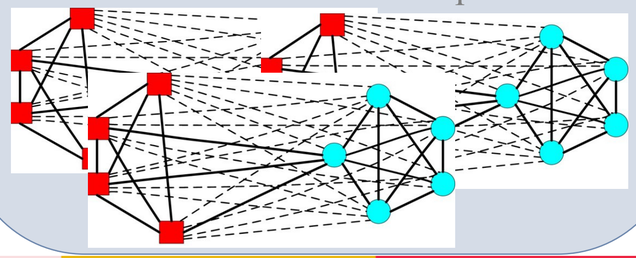
Multi-modal embedding in signed network



Frustration Cloud

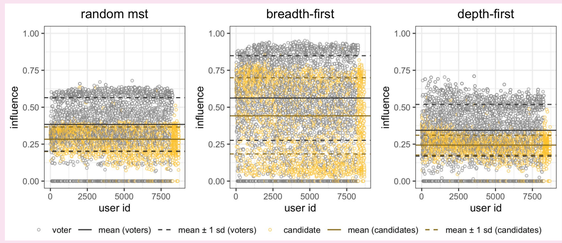


Status and Influence computation

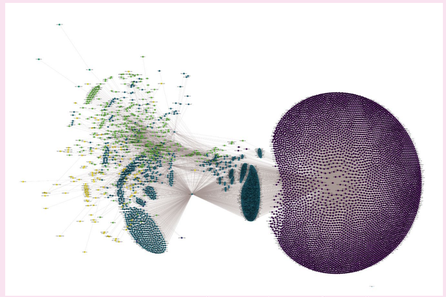


Output

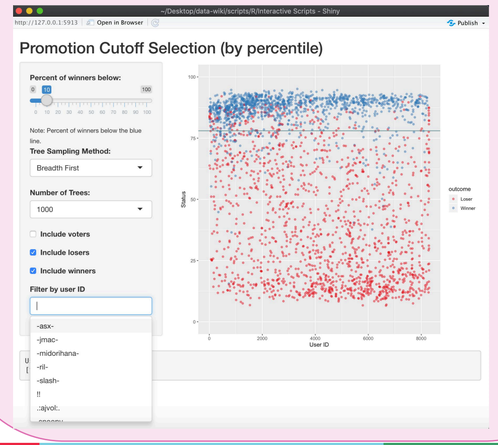
Modeling Analysis



Data Analysis



Interactive Analysis



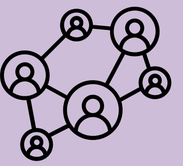
Outcome

Policy Bias Evaluation

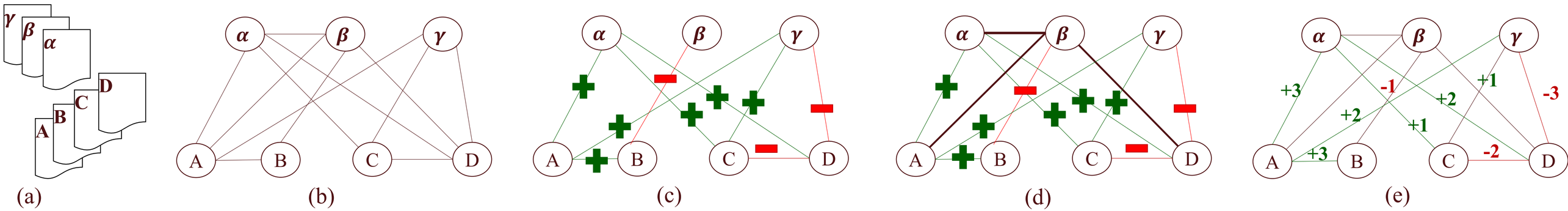
Outcome Review

Promotion Candidates

Communities



HOW DO WE MODEL UNSTRUCTURED DATA RELATIONSHIPS?

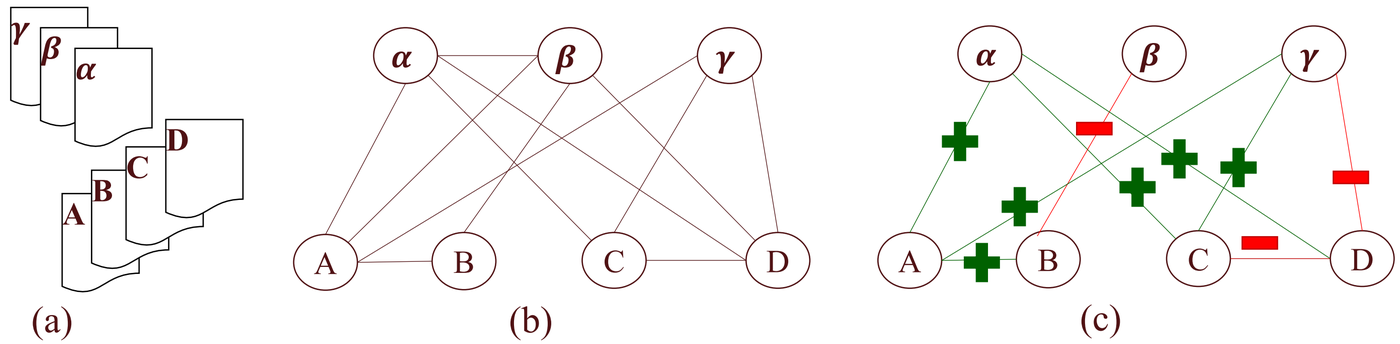


Unstructured data (a) and their representations: (b) unsigned graph for relationship; (c) signed graph for attitude; (d) merged (b) and (c); and (d) normalized weighted graph

Unstructured data does not conform to pre-defined data model or it is not organized e.g. Tweets, Health records, Open ended survey feedback, recommendation

Unstructured data need rich graph representation that unsinged GNN does not capture well

SIGNED GRAPH FROM REAL DATA

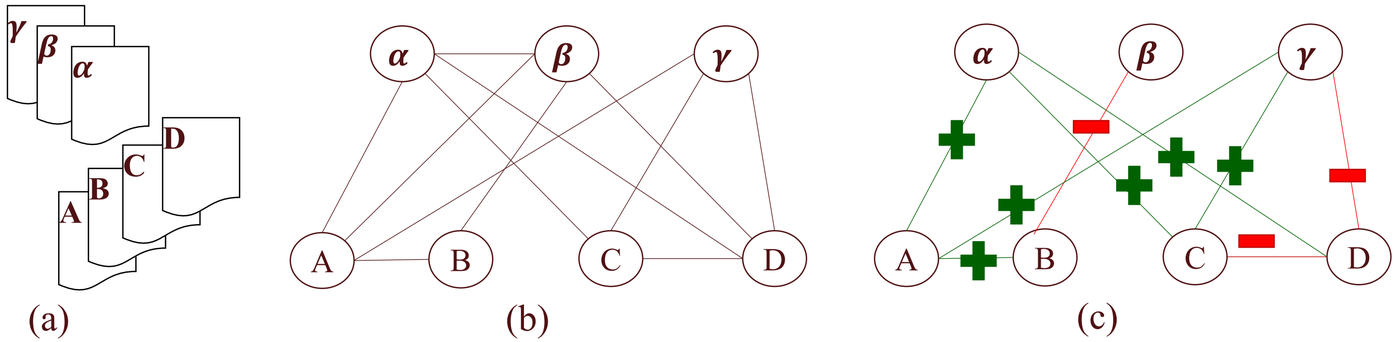


Signed graphs offer the binary sentiment relationship model

State-of-the-art in unsigned homogeneous graph tackles trillions of edges and billions of nodes (KDD ‘22) while signed graph benchmarking is at thousands of nodes and hundreds of thousands of edges (SDM ‘22).

- small in size and number – 12yo benchmark
- too similar in topology to support the research progress of signed graph analysis for real data

SIGNED GRAPH STATE OF THE ART



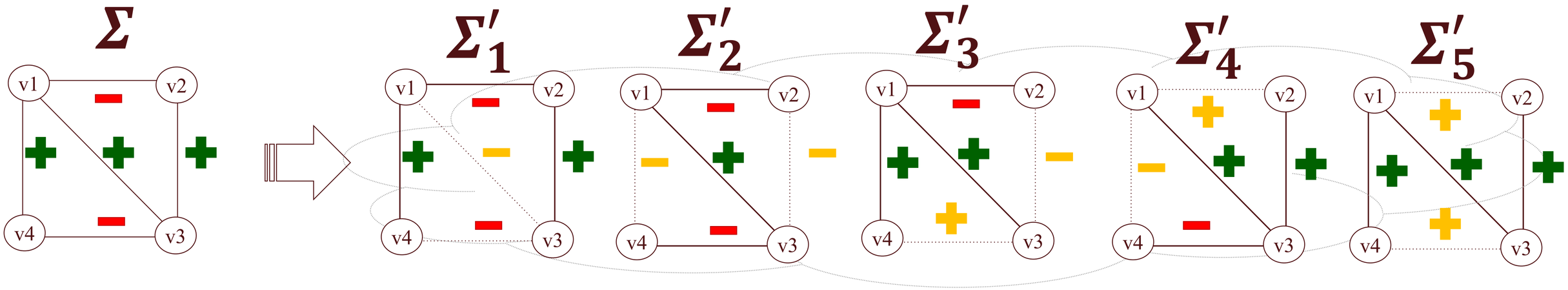
Signed graphs SOTA relies on spectral methods or GNN

- Spectral Methods suffer from eigenvector poisoning and scalability issues (Journal of Complex Networks, June 2022)
  - Small world and density assumption
  - Prohibitive complexity for real networks
- GNN advances for specific dataset and measure only – highly biased (KDD ‘22)
- Small and sparse benchmarks - **Advances in silos**

OUR SOLUTION: SCALABLE GRAPH BALANCING

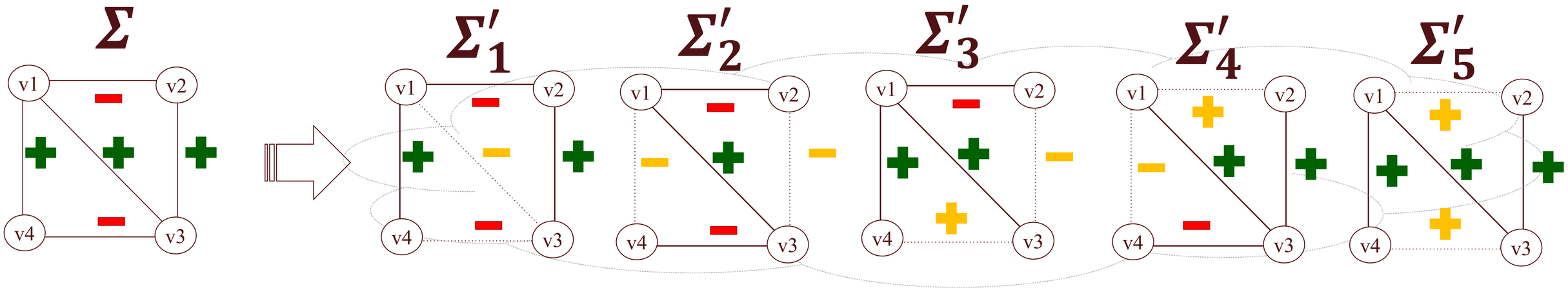
BALANCED STATES OF THE SIGNED GRAPH (DM&KD 2021)

- Balanced graph: signed graph where each of its cycles includes an even number of negative edges.
- Sociologists, psychologists, physicists, and control theorists are interested in the smallest number of edges whose sign can be changes so that the graph can be converted to balanced graph.
- Multiple options: balanced states



BALANCED STATES OF THE SIGNED GRAPH (DM&KD 2021)

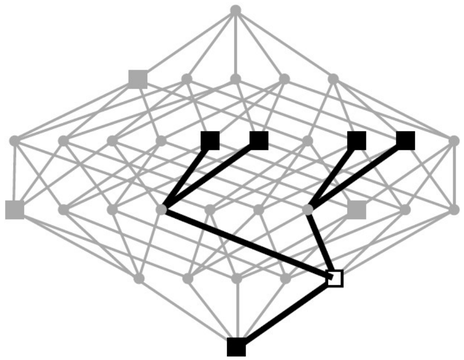
- The frustration index determines the distance of a network from a state of total structural balance.
- There is more that one way to achieve total structural balance by switching signs of minimum number of edges
- Frustration cloud: a set of all nearest balanced stated of the graph Σ

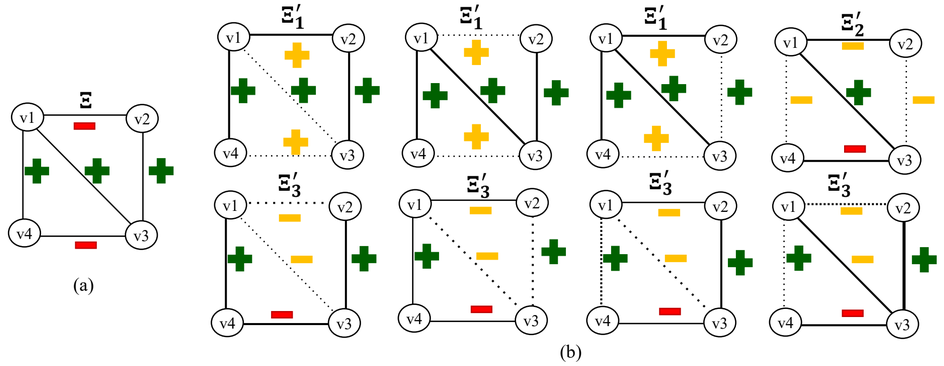


BALANCED STATES OF THE SIGNED GRAPH (DM&KD 2021)

- **Frustration cloud: a set of all nearest balanced stated of the graph Σ**
- **Balanced states cannot be easily found**

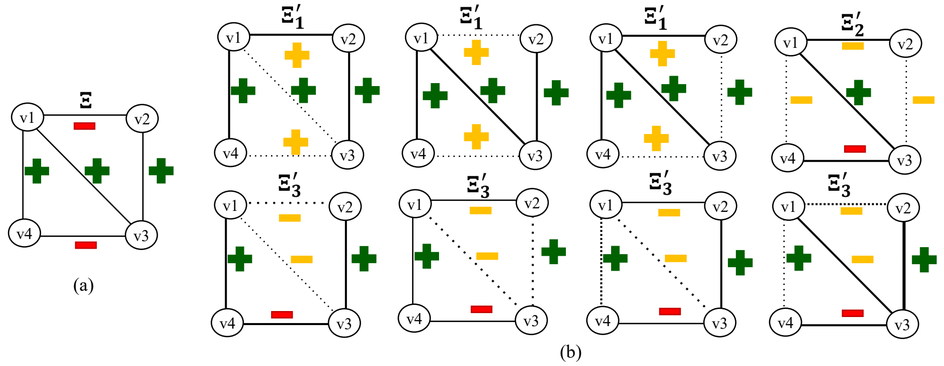
OUR PROPOSAL: TREE-BASED SAMPLING METHOD

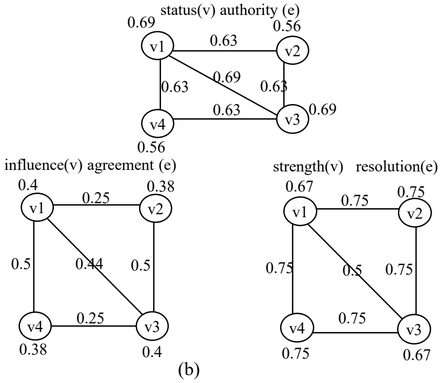




CONSENSUS FEATURES OF THE SIGNED GRAPH (DM&KD 2021,ACM SAC 2022)

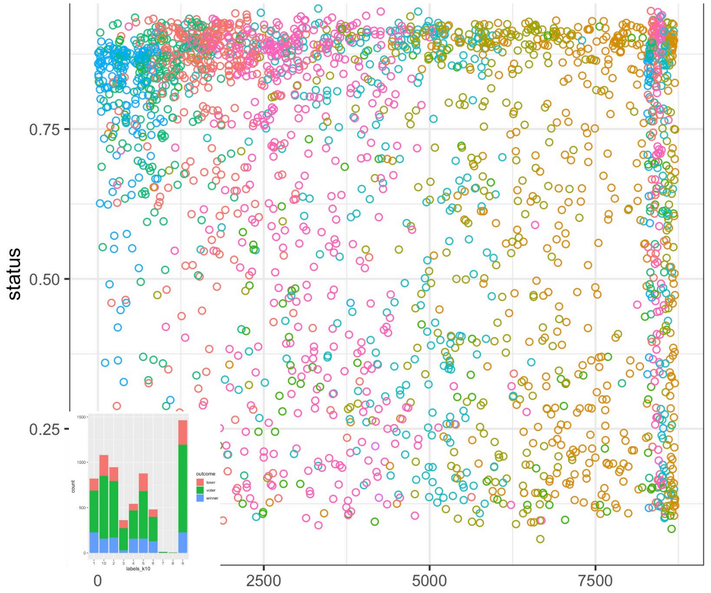
- **Characterize vertices using frustration cloud**
- **Consensus Space construction (In Submission)**



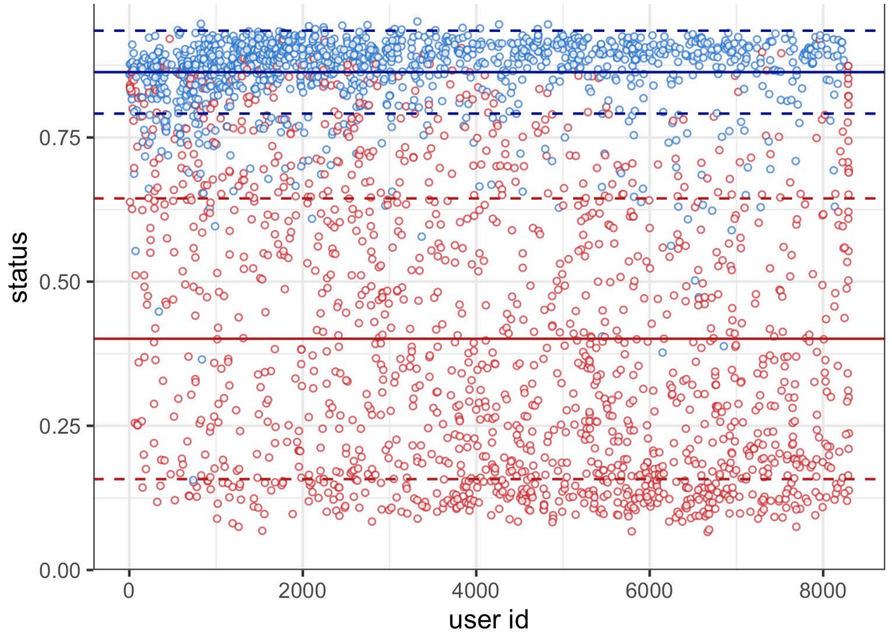


TREE-SAMPLING METHOF (DM&KD 2021)

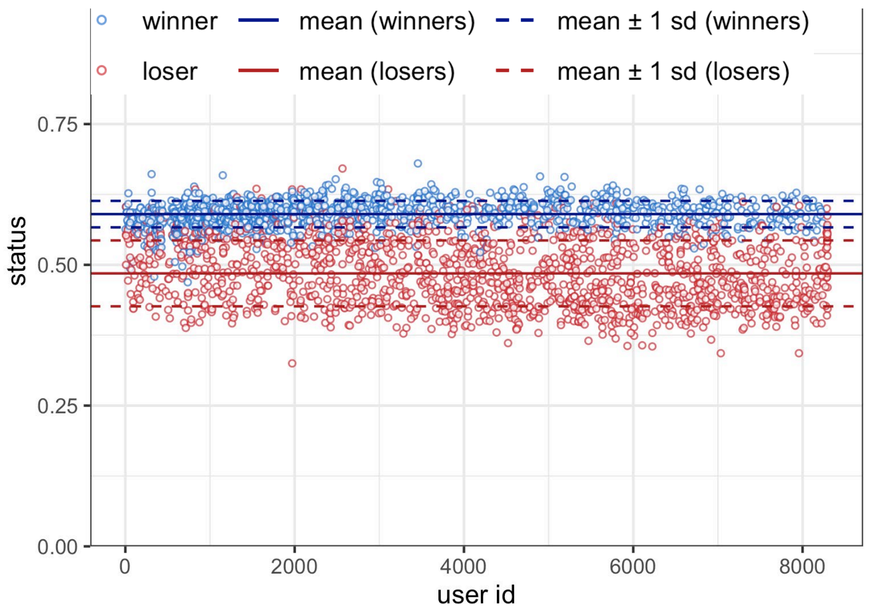
➢ **Breath-first search provides the highest resolution of the nearest balanced states**



(a) spectral clustering

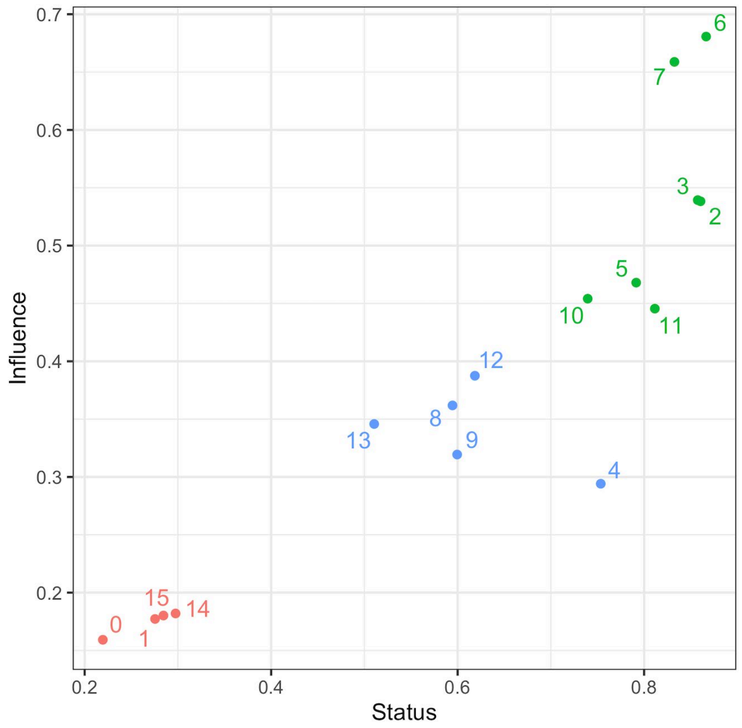


(b) BFS: outcome

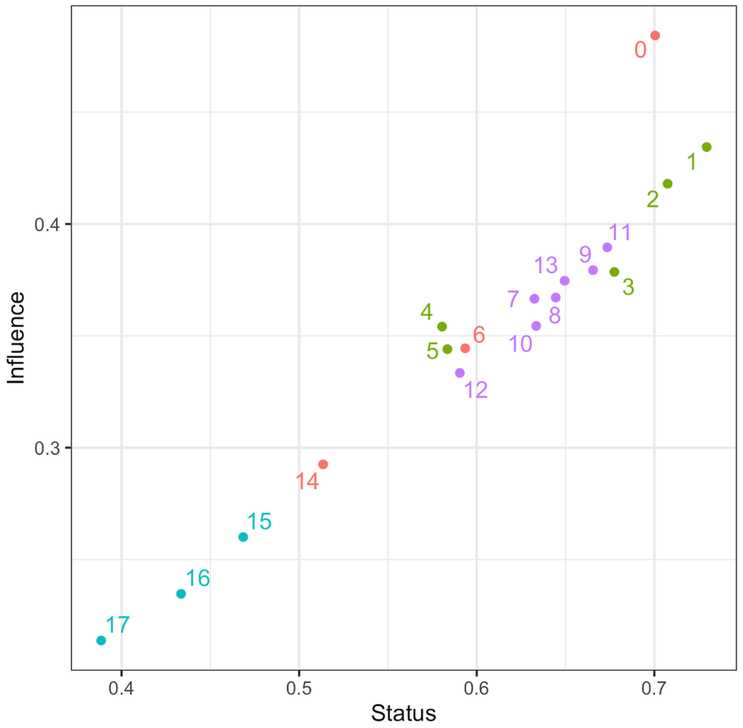


(c) random outcome

CONSENSUS BASED CLUSTERING (ACM-SAC '22)



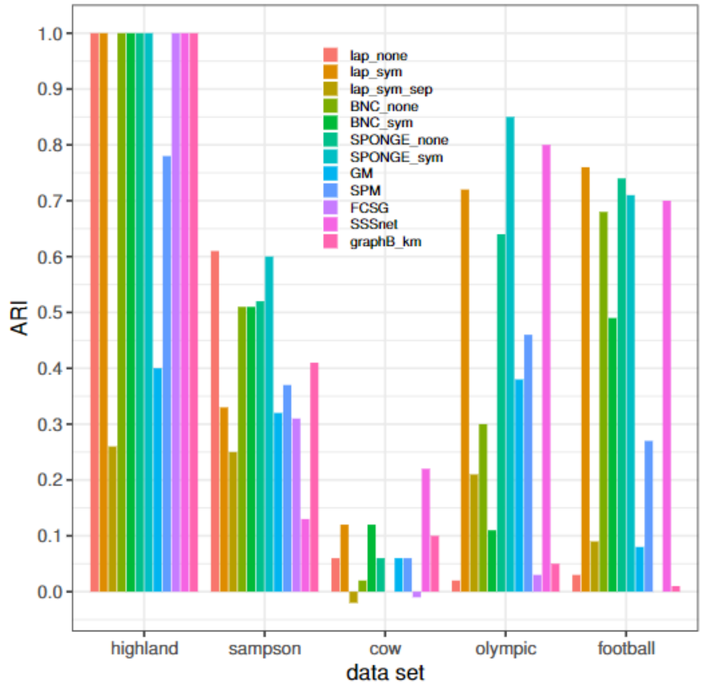
Highland Consensus based clustering of vertices show the strong correlation of cluster discoverability to GT

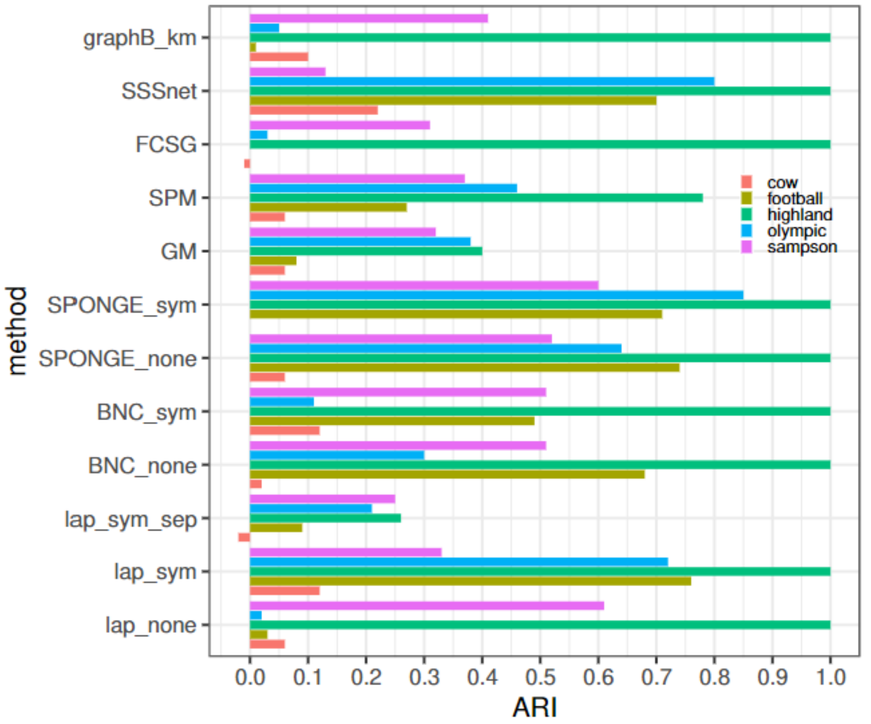


Sampson Consensus based clustering of vertices exposes the disconnect between GT and graph representation

SIGNED GRAPH CLUSTERING SURVEY (Journal of Complex Networks '22)

12 methods 5 labeled datasets





SIGNED GRAPH CLUSTERING SURVEY (Journal of Complex Networks '22)

5 methods 4 unlabeled datasets: scalability, runtime, trivial class recovery.

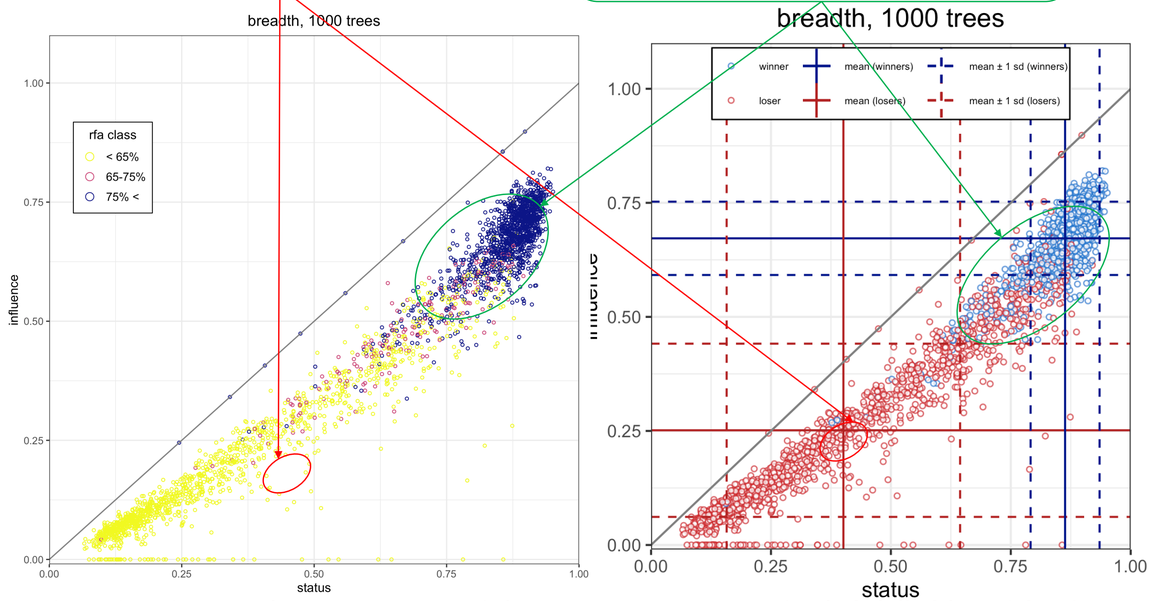
Most comprehensive and unbiased real benchmark to date

dataset	k	Laplacian		Balanced Cuts		SPONGE		FCSG		graphB		graphB		
		pos_in	neg_out	pos_in	neg_out	pos_in	neg_out	pos_in	neg_out	pos_in	neg_out	k	pos_in	neg_out
cow	3	**0.99**	**0.89**	1.0	0.55	1.0	0.3	0.84	0.25	0.98	0.58	3	0.98	0.58
wiki	30	**0.63**	**0.67**	0.9	0.17	0.59	0.71	0.49	0.52	0.05	0.96	4	0.29	0.73
slashdot	100	1.0	0.0	0.96	0.19	1.0	0.0	N/A	N/A	0.02	0.98	10	**0.22**	**0.78**
Epinions	100	1.0	0.0	**0.96**	**0.19**	1.0	0.0	N/A	N/A	0.03	0.97	10	0.13	0.88

DataLab12.github.com/graphB

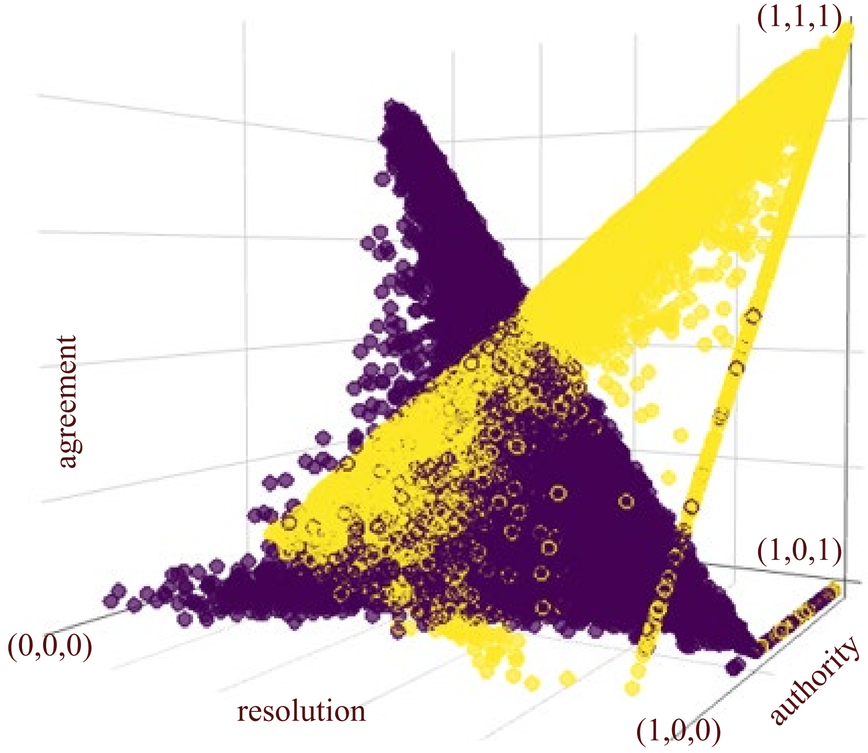
Users with los status and low influence and RfA in 65-75 % range elected

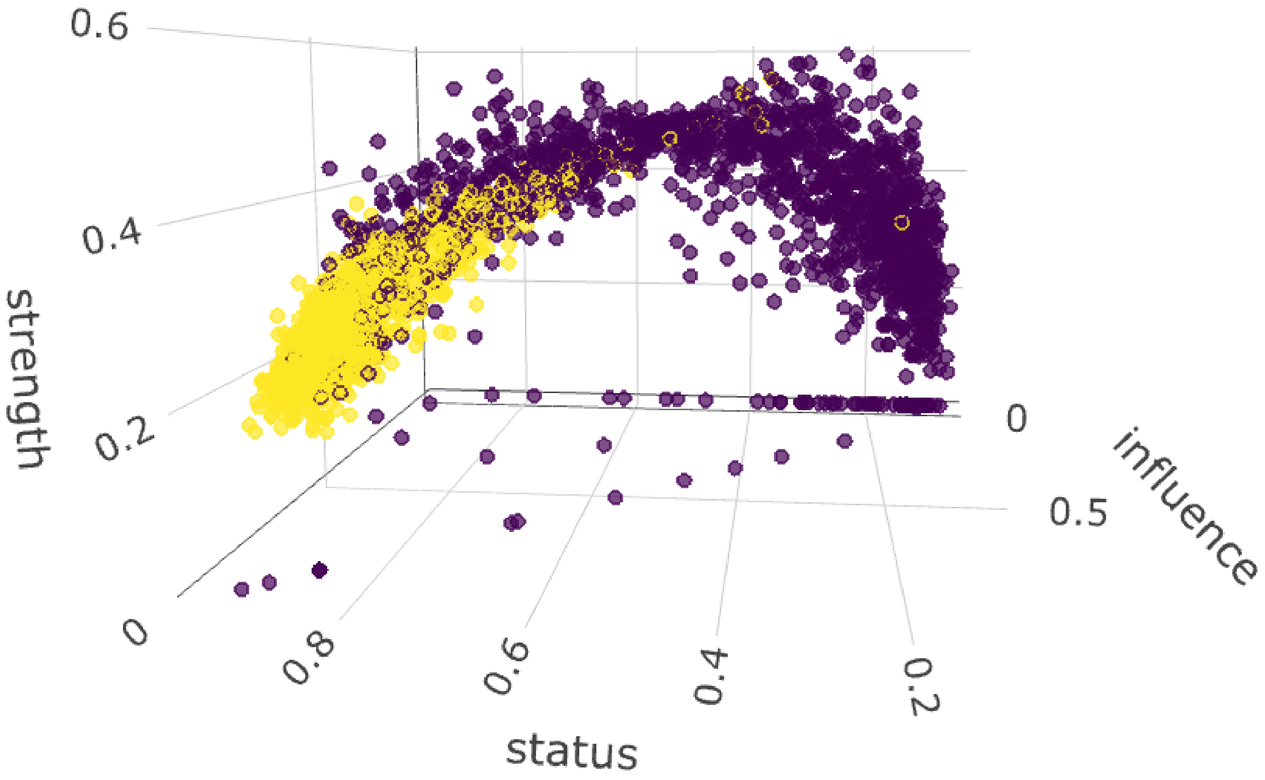
Users with high status and high influence and RfA in 65-75 % range not elected



Wikipedia election (over 7000 people) majority voting results (RfA) (left) and final outcome (right) wrt to status and influence measure the tool introduces. The tool flagged spam users, privileged users, narrow domain users and all anomalies in the process using simple rules:

BIAS DISCOVERY





DataLab12.github.com/graphB

Joint work w Prof. Rusnak, Math dept

Funded by TXST startup 2019 – 2021, and CHERR (2021-2022)

- Novel algorithm for signed graph analysis using balancing theory.
- Accurately models the alliance network
- Provides discriminant unbiased features for community discovery
- Successfully predicts administrator election outcome consistent with real election outcomes
- Balance theory answer to spectral clustering issues
- Scalable implementation w Dr. Burtcher's team (graphB+) to apply to Amazon data (SC '21)

Thank you! - jtesic@txstate.edu

- Computers will always do literally, exactly what you tell them to