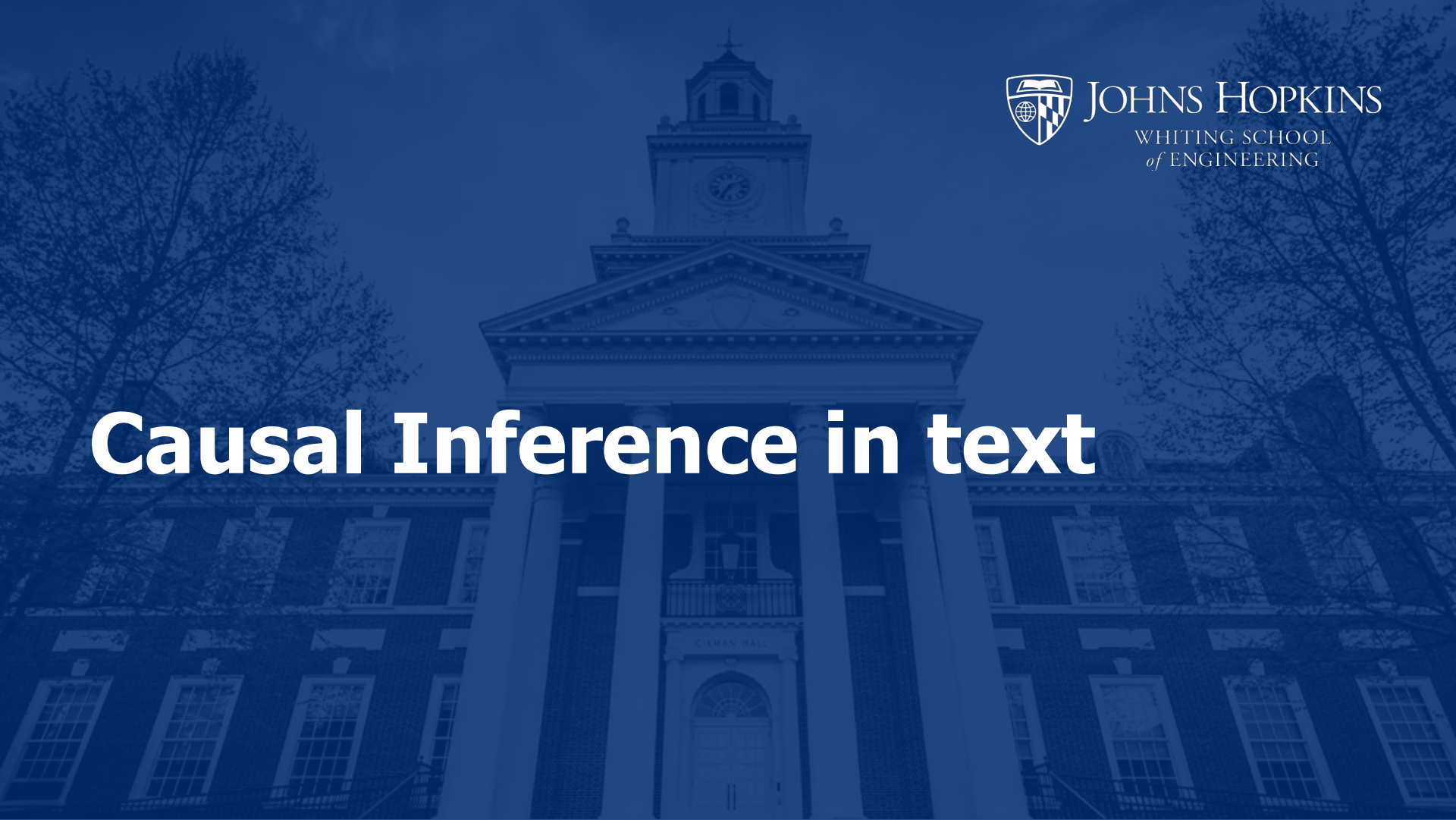
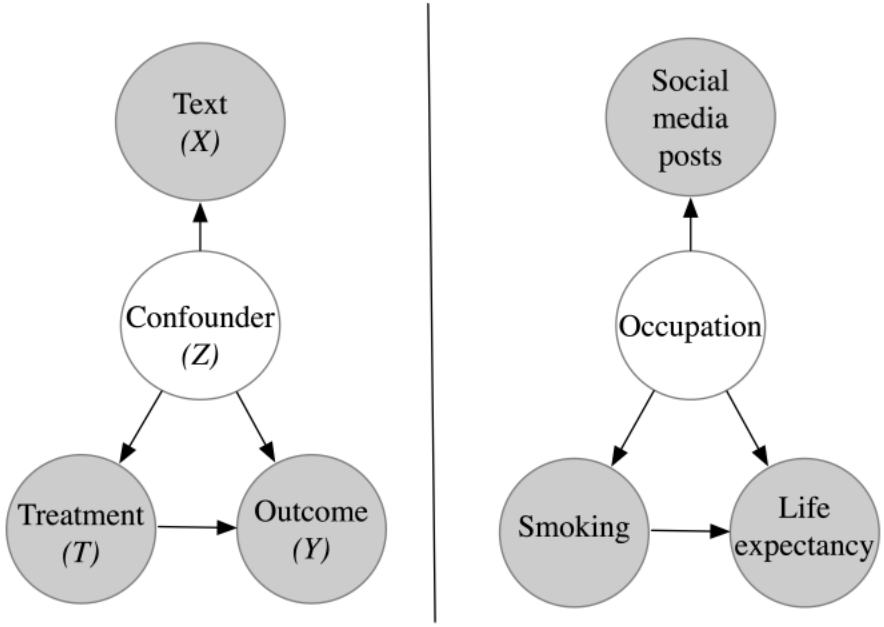


Causal Inference in text

Text as confounders

- Text data could either:
- (a) serve as a surrogate for potential confounders
- (b) the language of text itself could be a confounder

Example: the linguistic content of social media posts (confounder) could influence censorship (treatment) and future posting rates (outcome)



Keith, Katherine, David Jensen, and Brendan O'Connor. "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates." ACL. 2020.

Two similar approaches

- Topic Inverse Regression Matching
  - Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. "Adjusting for confounding with text matching." American Journal of Political Science 64.4 (2020): 887-903.
- "Causally sufficient" embeddings
  - Veitch, Victor, Dhanya Sridhar, and David Blei. "Adapting text embeddings for causal inference." Conference on Uncertainty in Artificial Intelligence. PMLR, 2020.

A different method: develop "causally sufficient" text embeddings

- Text is high dimensional and data is finite: difficult to fit models directly to text
- Instead, "reduce the text to a low-dimensional representation that suffices for causal identification and enables efficient estimation from finite data."

- Two key ideas:
  - Supervised dimensionality reduction: we don't need the full text, causal inference only requires the parts of text that are predictive of the treatment and outcome
  - Efficient language modeling: design representations of text to dispose of "linguistically irrelevant information", presumed to also be "causally irrelevant"

Veitch, Victor, Dhanya Sridhar, and David Blei. "Adapting text embeddings for causal inference." Conference on Uncertainty in Artificial Intelligence. PMLR, 2020.

General approach: develop "causally sufficient" text embeddings

- Start with a neural NLP model (BERT language model) and modify it to produce 3 outputs:
  - 1) document-level embeddings
  - 2) a map from the embeddings to treatment probability
  - 3) a map from the embeddings to expected outcomes for the treated and untreated
  - [(2) and (3) are small added neural networks on the original model]

- [They also do a variant based on a topic model]

General approach: develop "causally sufficient" text embeddings

- Train model to predict outcome, treatment, and with language-modeling objective (e.g. to learn meaningful text representations)

$$L(\mathbf{w}_i;\xi,\gamma) = \big(y_i - \tilde{Q}(t_i,\lambda_i;\gamma)\big)^2 \longrightarrow \text{Outcome}$$
$$+ \,\text{CrossEnt}\big(t_i,\tilde{g}(\lambda_i;\gamma)\big) \longrightarrow \text{Treatment}$$
$$+ \,L_{\text{U}}(\mathbf{w}_i;\xi,\gamma). \longrightarrow \text{Language modeling}$$

- To compute average treatment effect, plug estimated embeddings, propensity scores, and conditional outcomes into a downstream estimator

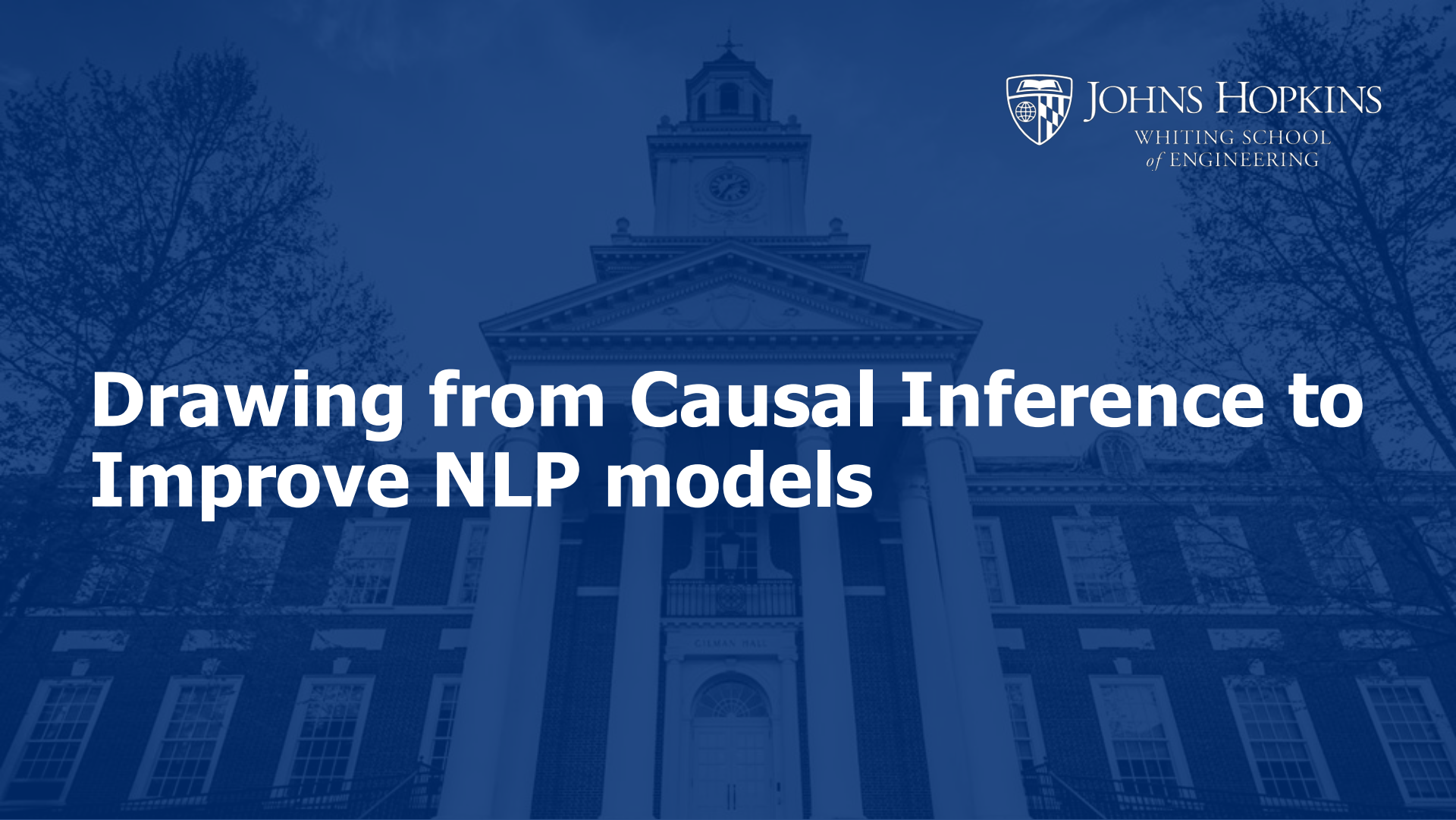
Evaluation

- Two settings:
  - Peer-reviewed journal articles: Causal effect of including a theorem on paper acceptance.
    - Treatment: the word "theorem" occurs in the paper
    - Confounder: article abstract (subject of the paper)
    - Outcome: accept/reject
  - Effect of gender on Reddit popularity
    - Treatment: "male" label
    - Mediator: Post text (topic or style)
    - Outcome: Popularity score

How can we use this data for *evaluation* rather than analysis?

Evaluations

- Simulated data:
  - Use real confounders and treatments
  - Simulate outcomes (so we know the "true" causal effect)
- Their findings:
  - 1) Yes, language modeling helps recover simulated effects
  - 2) Yes, supervised dimensionality helps
  - 3) Their proposed models C-BERT and C-ATM outperform alternatives



Drawing from Causal Inference to Improve NLP models

Drawing from Causal Inference to Improve NLP models

- ML in general typically captures associates, not causal effects
- Models are prone to overfitting, exploit spurious correlations in the data
  - E.g. train a model to identify photos of dogs from cats; Model learns that dogs always have collars

→ "DOG"

Drawing from Causal Inference to Improve NLP models

- ML in general typically captures associates, not causal effects
- Models are prone to overfitting, exploit spurious correlations in the data
  - E.g. train a model to identify photos of dogs from cats; Model learns that dogs always have collars

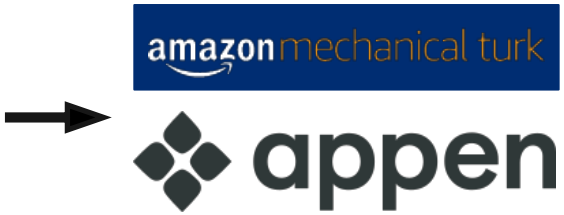
- Maybe by drawing from causal inference we can train models to ignore these spurious correlations, especially for tasks where it's hard to collect good training data
- Case study: drawing from causal inference to detect *subtle gender bias*

Field, Anjalie, and Yulia Tsvetkov. "Unsupervised Discovery of Implicit Gender Bias." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

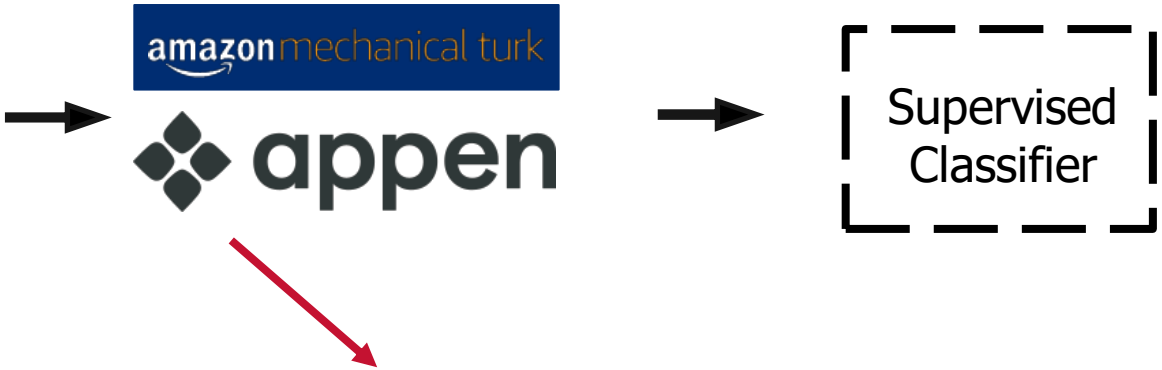
Need to develop new models

- Our goal: detect subtle gender biases like microaggressions, objectifications, and condescension in 2nd-person text
  - "Oh, you work at an office? I bet you're a secretary"
  - "Total tangent I know, but you're gorgeous"
- Current classifiers that detect hate speech, offensive language, or negative sentiment cannot detect these comments
- [Note: focus on binary gender]

Naive Approach: Supervised Classification



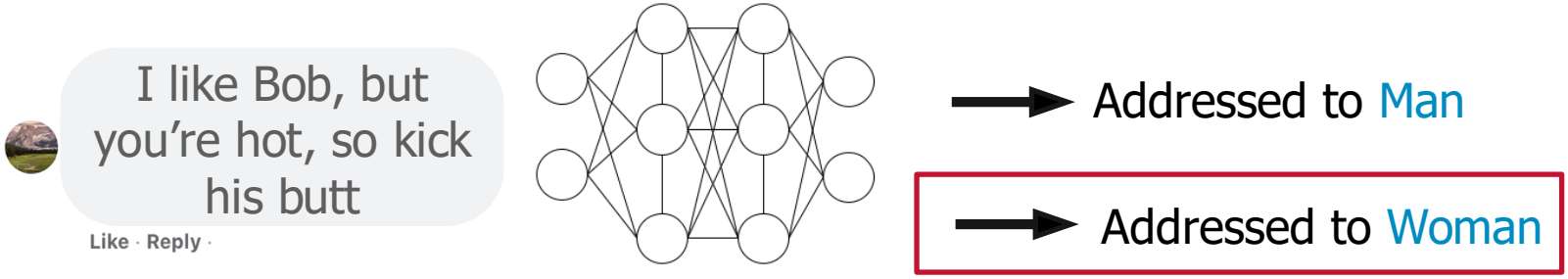
Naive Approach: Supervised Classification



Problem: Biases are *subtle*, *implicit*, and *context-dependent*

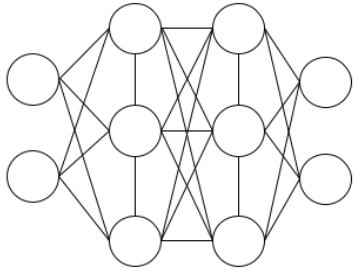
Proposed approach: Comments contain gender bias if they are highly predictive of gender

- Train a classifier that predicts the gender of the person the text is addressed to
- If the classifier makes a prediction with high confidence, the text likely contains bias



If a comment is very likely to be addressed to a woman, and is very unlikely to be addressed to a man, it probably contains gender bias.

Challenge: Text main contain *confounds* that are predictive of gender, but not indicative of gender bias



→ Addressed to Woman

→ Addressed to Woman

→ Addressed to Woman

→ Addressed to Woman

Challenge: Text main contain *confounds* that are predictive of gender, but not indicative of gender bias

- Overtly gendered words
- Preceding context in the conversation
- Traits of people (other than gender) in the conversation

Saturday is the 40th anniversary of **Title IX**…

Like · Reply ·

…a good way to celebrate Title IX, too!

Like · Reply ·

I'd vote for you if I lived in Massachusetts

Like · Reply ·

Bob and I join Bill Hemmer on America's Newsroom to discuss whether or not...

Like · Reply ·

I like Bob, but you're hot, so kick his butt

Like · Reply ·

Thanks so much Ma'am!

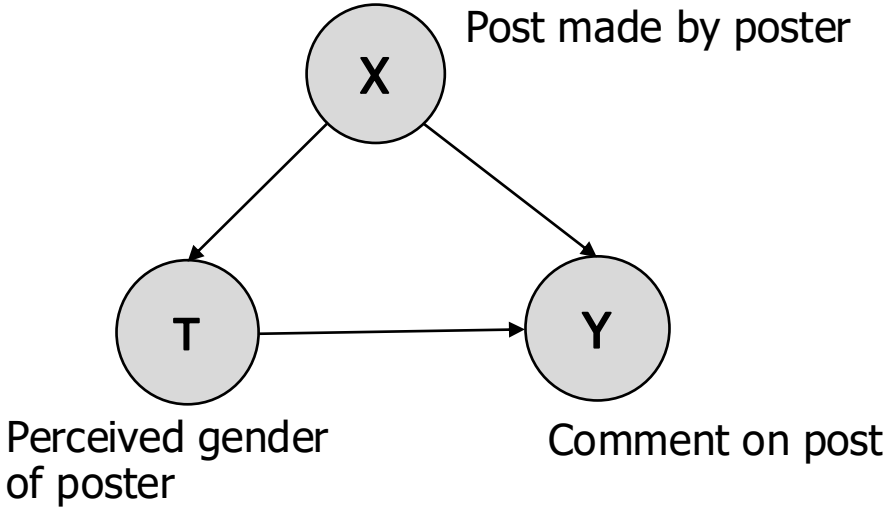
Like · Reply ·

A note on causal set-up

- We're not really doing causal inference: we are trying to build a classifier to detect microaggressions, not draw conclusions about the state of the world:
  - "confounds": spurious correlations in our data (not necessarily "confounders")

- Some of these factors that we don't want the model to learn are confounding variables

A note on causal set-up



[Note: we have text as an outcome and as a confounder]

Preceding context is an *observed* confounding variables

Writer_Gender: F

Saturday is the 40th anniversary of **Title IX**! I'm celebrating with a Sat morning run - ladies please respond below if you want to join

Like · Reply ·

Wish I could ! Already have plans for a bike ride and breakfast with some awesome ladies - a good way to celebrate **Title IX**, too!

Like · Reply ·

Would love to!

Like · Reply ·

Writer_Gender: M

Any deal with **Iran** — a nation that the United States cut off diplomatic ties with 35 years ago — must protect America's interests at home and abroad.

Like · Reply ·

Iran might be a free, democratic nation today, if not for decades of American interference.

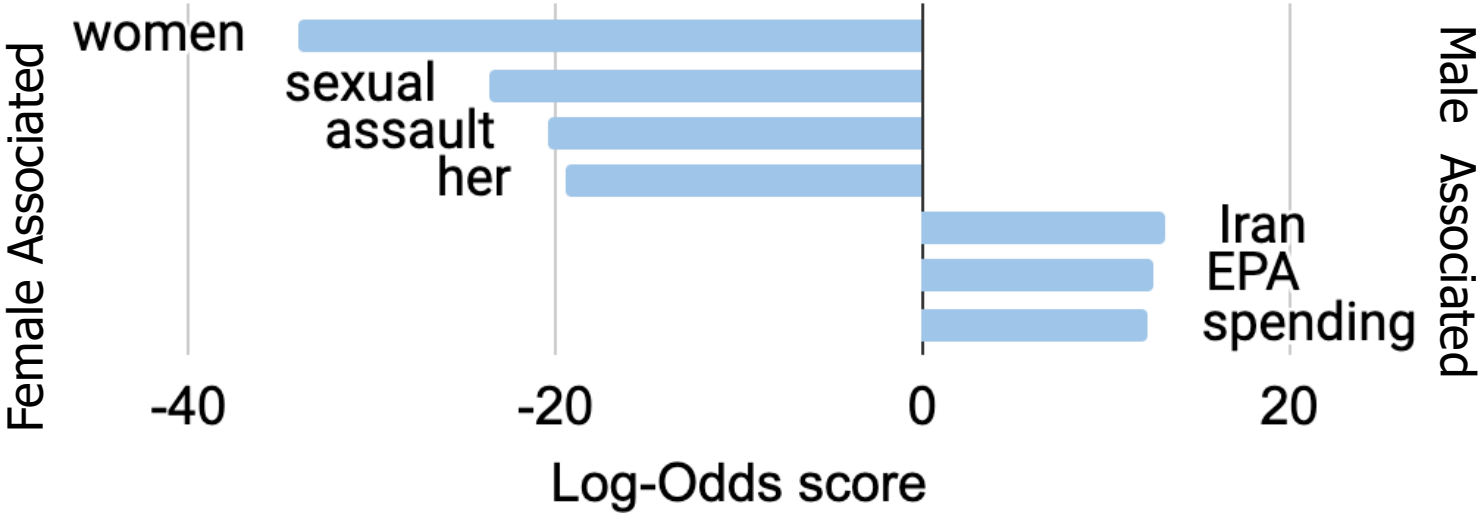
Like · Reply ·

That's for sure! Worst deal he could make! We can't trust **Iran** and America knows it !!!!!

Like · Reply ·

Key problem: Men and women post different content, which is reflected in their replies

Preceding context is an *observed* confounding variables



Propensity matching for *observed* confounding variables

Writer_Gender: F

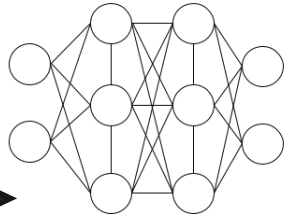
~~Saturday is the 40th anniversary of **Title IX**! I'm celebrating with a Sat morning run - ladies please respond below if you want to join~~

Writer_Gender: M

Any deal with **Iran** — a nation that the United States cut off diplomatic ties with 35 years ago — must protect America's interests at home and abroad.

Writer_Gender: F

My overriding concern is whether or not the agreement is in the national security interest of the United States. **Iran** must be blocked from proceeding any further towards developing a nuclear weapon.



Text classifier to predict WRITER_GENDER

$$|e_i - e_l| \geq c \forall l$$

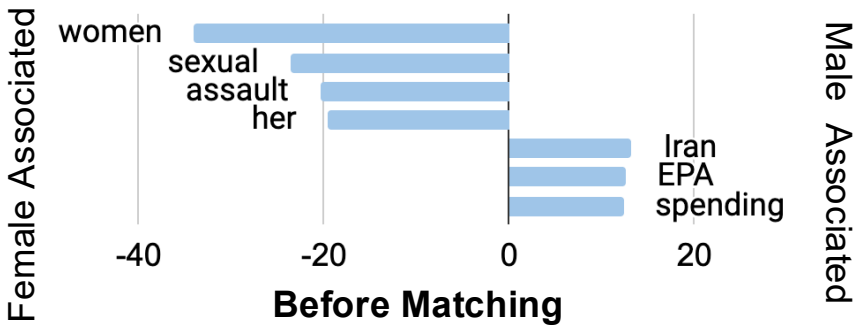
$$e_i = P(W.Gender_i = F | Post_i) \approx 0.91$$

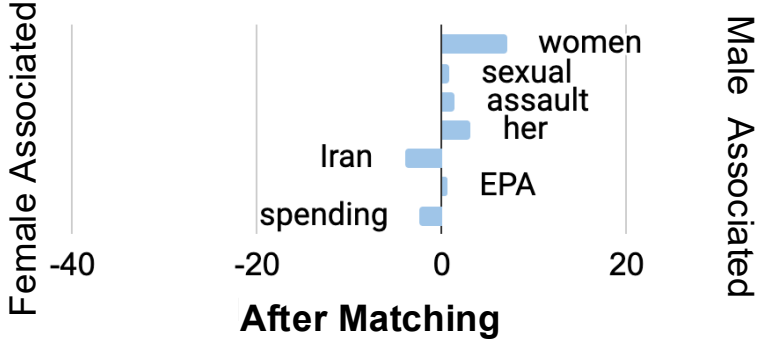
$$e_j = P(W.Gender_j = F | Post_j) \approx 0.33$$

$$e_k = P(W.Gender_k = F | Post_k) \approx 0.32$$

$$argmin_j |e_k - e_j|$$

Propensity matching for *observed* confounding variables

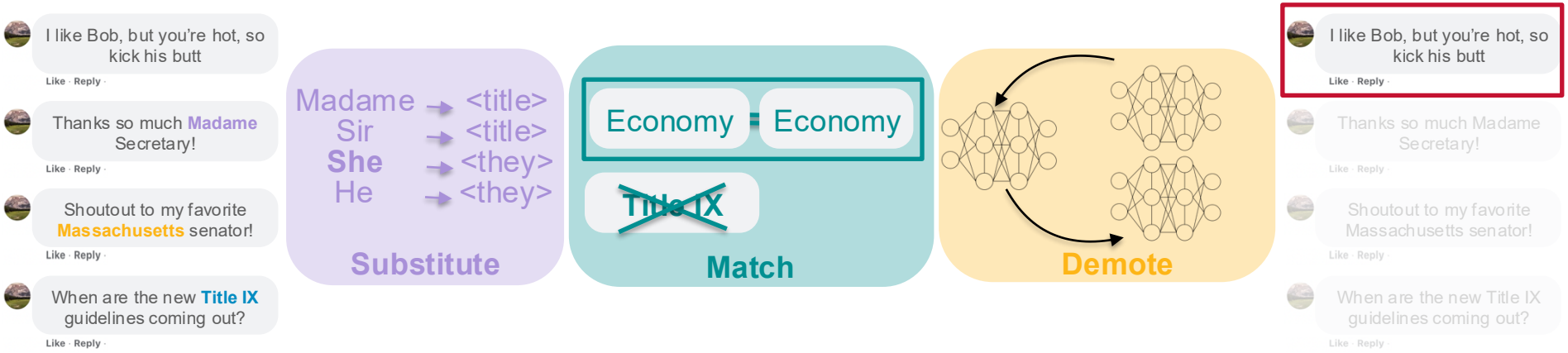




Propensity matching breaks associations between gender and context in the training data

Proposed Model: Comments contain bias if they are highly predictive of gender *despite confound control*

- Substitute overt indicators
- Balance observed confounders through propensity matching
- Demote latent confounders through adversarial training



Self-reported microaggressions

	Public Figs		Politicians	
	F1	**Acc.**	**F1**	**Acc.**
base	61.3	57.3	48.1	64.2
+demotion	**62.2**	57.9	53.7	61.5
+match	38.9	55.9	46.9	50.7
+match+dem.	50.9	57.0	**56.9**	49.9
Random	46.0	49.8	-	-
Class Random	42.1	48.3	-	-

- Models are not trained at all for this task; they are only trained for gender-of-addressee prediction, but they still perform better than chance

Findings: characteristics of bias against women politicians

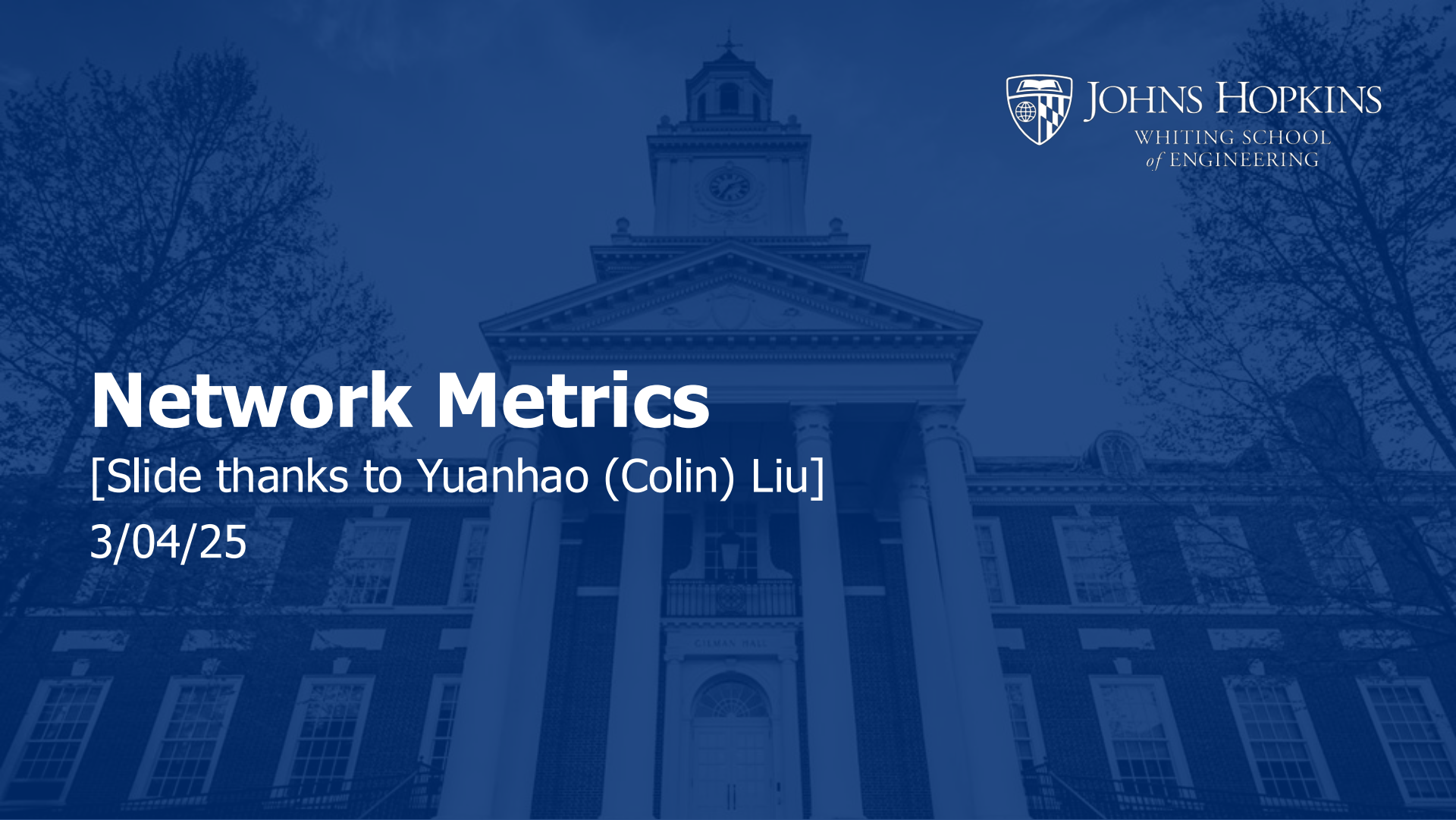
- Influential words:
  - Competence and domesticity
  - 'Force', 'situation', 'spouse', 'family', 'love'

- Examples:
  - "DINO I hope another real Democrat challenges you next election"
  - "I did not vote for you and have no clue why anyone should have. You do not belong in politics"

Findings: characteristics of bias against women

- Influential words:
  - Appearance and sexualization
  - ‘beautiful’, ‘love’,‘sexo’

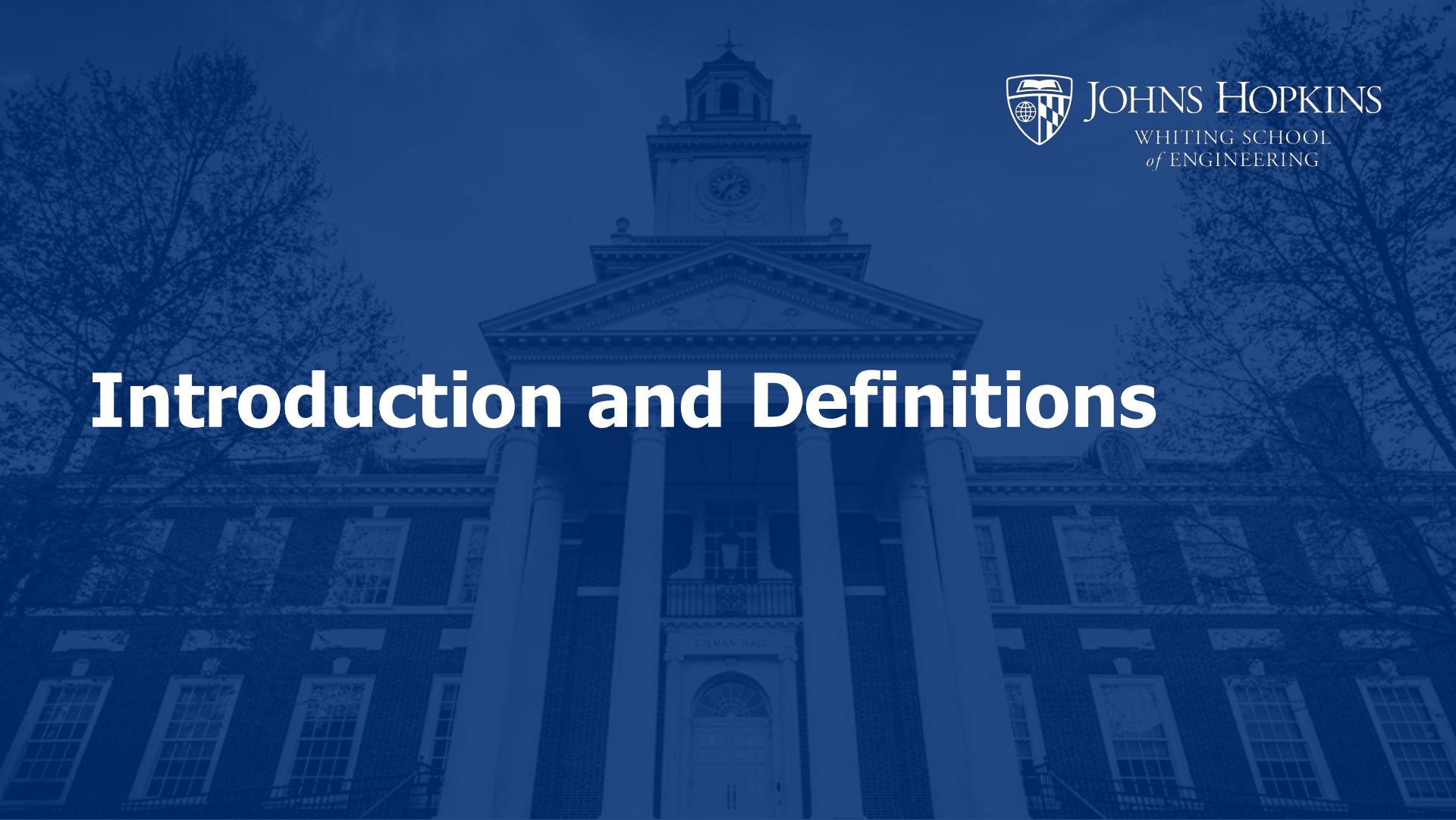
- Examples:
  - “Total tangent I know but, you’re gorgeous.”
  - “I like Bob, but you’re hot, so kick his butt.”



Network Metrics

[Slide thanks to Yuanhao (Colin) Liu]

3/04/25



Introduction and Definitions

Motivation: understand online "epidemic"

- Lies spread faster than the truth

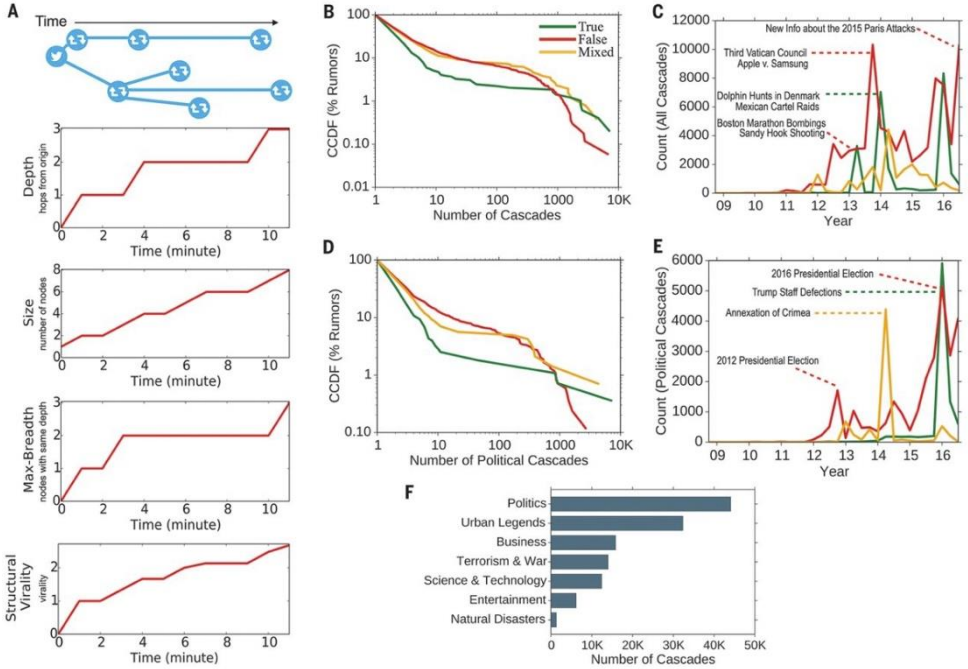


Fig. 1 Rumor cascades.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. science, 359(6380), 1146-1151.

Motivation: understand epidemic

- Sex Partner Network and HIV

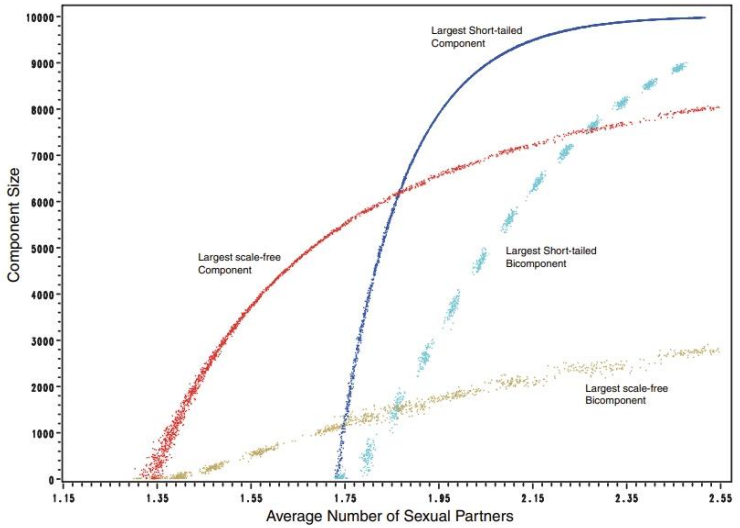


Fig. 3. Size of Largest component and bicomponent by average number of sexual partners for short-tailed and scale-free distributions. The curves plot the growth of the largest component and bicomponent as a function of the average degree, based on 100 simulations of a 10,000-node network at each degree setting. The red curve plots the analytic solution for the size of the giant component for the simulated networks with scale-free distributions, and the orange curve plots the largest bicomponent. The dark blue curve plots the analytic solution for the size of the largest component for the simulated low-degree networks, and the light blue curve plots the size of the largest bicomponent. The bicomponent curves are not continuous due to sampling.

Moody, J., Adams, J., & Morris, M. (2017). Epidemic potential by sexual activity distributions. Network science, 5(4), 461-475.

Motivation: how to succeed as individual

- Looking for a job? Making Weak Ties.
- Want to be influential? Try something new, but don't go too far.

The Strength of Weak Ties[1]

Mark S. Granovetter
Johns Hopkins University

Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

Atypical Combinations and Scientific Impact

Brian Uzzi,[1,2] Satyam Mukherjee,[1,2] Michael Stringer,[2,3] Ben Jones[1,4]*

Novelty is an essential feature of creative ideas, yet the building blocks of new ideas are often embodied in existing knowledge. From this perspective, balancing atypical knowledge with conventional knowledge may be critical to the link between innovativeness and impact. Our analysis of 17.9 million papers spanning all scientific fields suggests that science follows a nearly universal pattern: The highest-impact science is primarily grounded in exceptionally conventional combinations of prior work yet simultaneously features an intrusion of unusual combinations. Papers of this type were twice as likely to be highly cited works. Novel combinations of prior work are rare, yet teams are 37.7% more likely than solo authors to insert novel combinations into familiar knowledge domains.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, *78*(6), 1360-1380.
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. Science, 342(6157), 468-472.

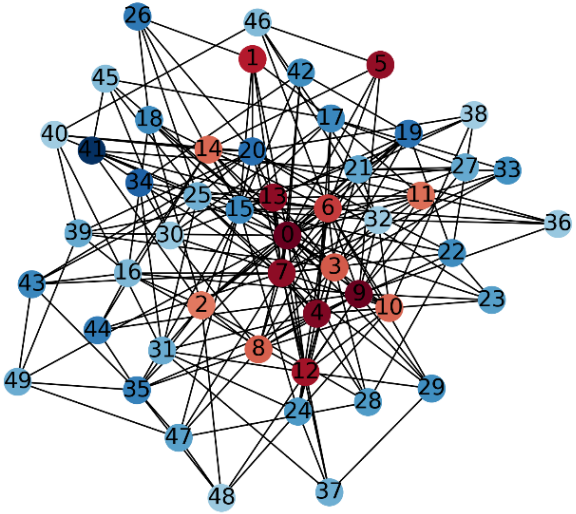
How might we represent network?

Represent connections between vertices/nodes

- Vertex: a node of the graph
- Edge: a link between two vertices

A graph consists of a set of nodes and a set of edges

- $G(V, E)$



Graph Data: Adjacency Matrix

- The matrix of vertices connections

Encode in a symmetric matrix (for undirected network)

$(n \times n)\ matrix\ A$

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

The adjacency matrix has elements

$$a_{ij} = \begin{cases} 1 & if\ i\ and\ j\ are\ connected \\ 0 & otherwise \end{cases}$$

	Mark	Peter	Bob	Jill	Aaron
Mark	0	1	0	1	0
Peter	1	0	1	0	1
Bob	0	1	0	1	0
Jill	1	0	1	0	1
Aaron	0	1	0	1	0

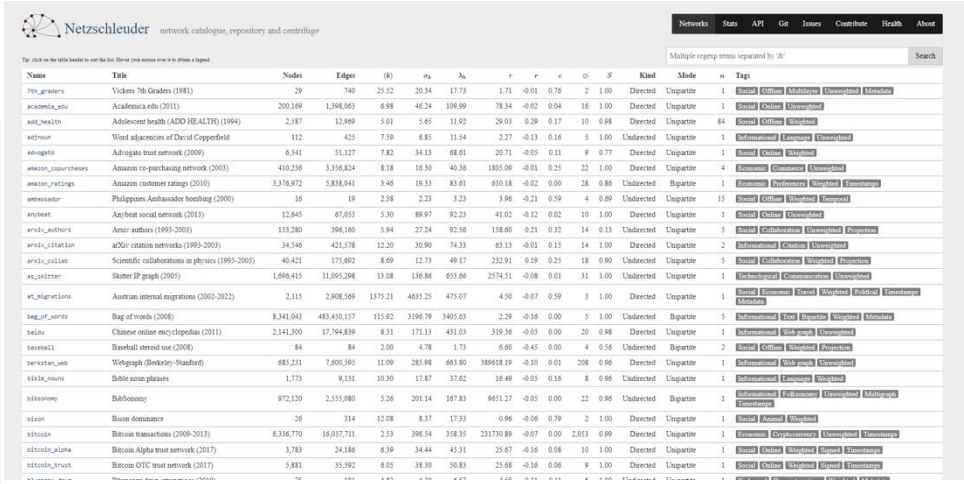
Example from: https://bookdown.org/markhoff/social_network_analysis/understanding-network-data-structures.html

Graph Data: Edge Lists

- Two-column matrices that directly indicate how vertices are connected

	Mark	Peter	Bob	Jill	Aaron
Mark	0	1	0	1	0
Peter	1	0	1	0	1
Bob	0	1	0	1	0
Jill	1	0	1	0	1
Aaron	0	1	0	1	0

	PersonA	PersonB
1	Mark	Peter
2	Mark	Jill
3	Peter	Bob
4	Peter	Aaron
5	Bob	Jill
6	Jill	Aaron



Example from: https://bookdown.org/markhoff/social_network_analysis/understanding-network-data-structures.html
Collections of Social Network Datasets: https://networks.skewed.de/

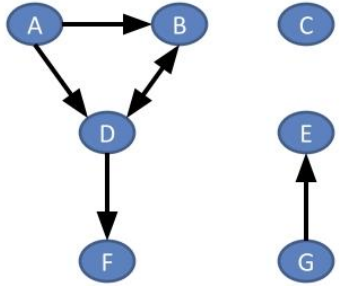
Types of Edges

- Directed vs. undirected

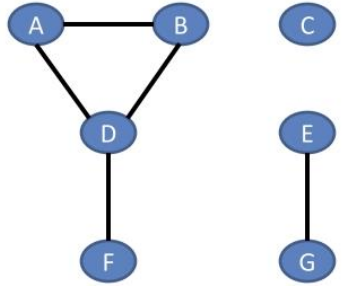
Directed & undirected

- Communication vs. friendship networks

twitter



facebook



Directed sociomatrix

	A	B	C	D	E	F	G
A	-	1	0	1	0	0	0
B	0	-	0	1	0	0	0
C	0	0	-	0	0	0	0
D	0	1	0	-	0	1	0
E	0	0	0	0	-	0	0
F	0	0	0	0	0	-	0
G	0	0	0	0	1	0	-

Unsymmetric

Symmetric

Unsymmetric

Undirected sociomatrix

	A	B	C	D	E	F	G
A	-	1	0	1	0	0	0
B	1	-	0	1	0	0	0
C	0	0	-	0	0	0	0
D	1	1	0	-	0	1	0
E	0	0	0	0	-	0	1
F	0	0	0	1	0	-	0
G	0	0	0	0	1	0	-

Symmetric

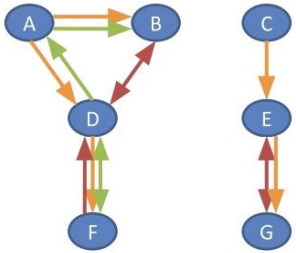
Symmetric

Symmetric

Types of Edges

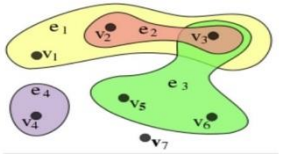
- Weighted vs. unweighted
- Multiplex

Organizations: authority, trust, & friendship

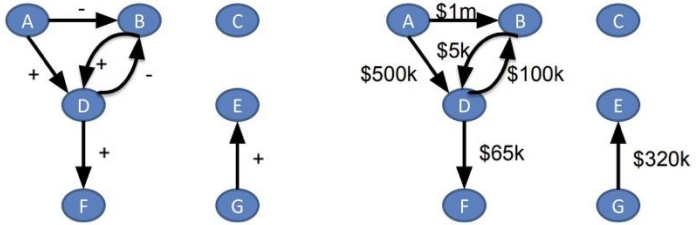


Hypergraph Incidence Matrix

	e_1	e_2	e_3	e_4
v_1	1	0	0	0
v_2	1	1	0	0
v_3	1	1	1	0
v_4	0	0	0	1
v_5	0	0	1	0
v_6	0	0	1	0
v_7	0	0	0	0

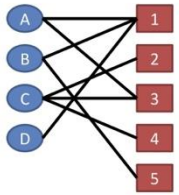


- Affect in a sorority vs. campaign financing



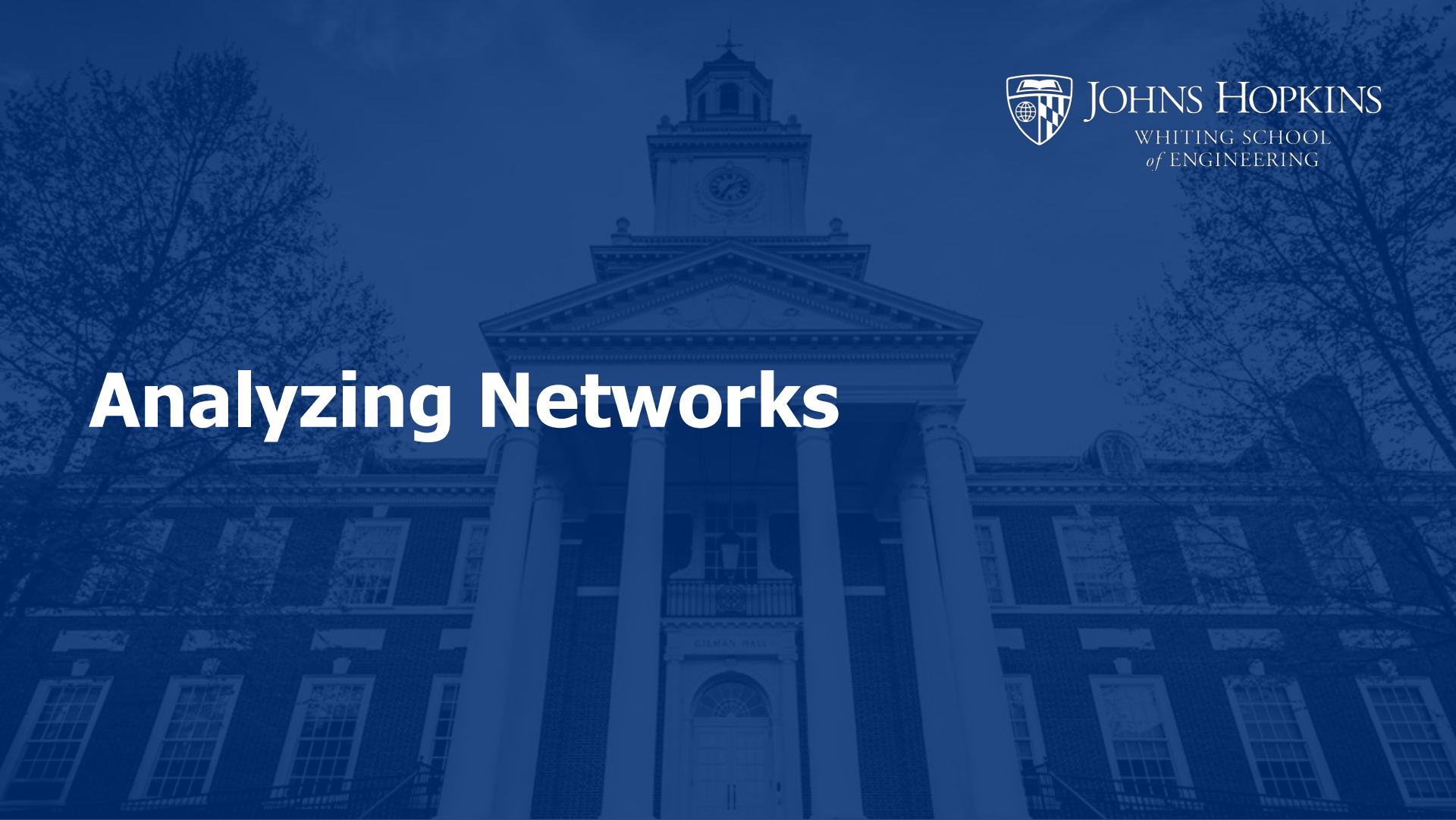
Bipartite sociomatrix

	1	2	3	4	5
A	1	0	1	0	0
B	1	0	0	0	1
C	0	1	1	1	0
D	1	0	0	0	0



Example from: https://sonic.northwestern.edu/

Example of hypergraph: Lungeanu, A., Carter, D. R., DeChurch, L. A., & Contractor, N. S. (2021). How team interlock ecosystems shape the assembly of scientific teams: A hypergraph approach. In Computational Methods for Communication Science (pp. 95-119). Routledge.



Analyzing Networks

Basic Metrics

- We can calculate many metrics directly from the network structure
- Example: Network density
  - Network density: $\frac{Number\ of\ edges}{Number\ of\ possible\ edges}$
  - For a directed unweighted network with n nodes, the max number of possible edges is: $n(n-1)$

Americans are becoming more isolated

Table 3. Structural Characteristics of Core Discussion Networks

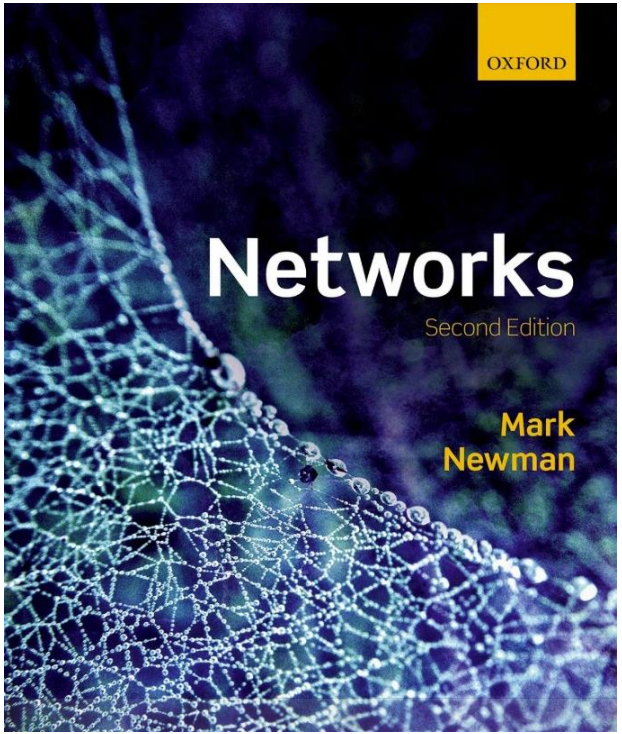
	1985 (N = 1,167[a])	2004 (N = 788[b])
Network Density		
<.25	9.9%	7.3%
.25–.49	18.5%	11.8%
.50–.74	37.9%	39.5%
>.74	33.7%	41.4%
Mean	.60	.66
SD	.33	.33
Mean Frequency of Contact (days per year)		
6–12	3.7%	3.0%
>12–52	15.3%	10.6%
>52–365	81.0%	86.4%
Mean	208.92	243.81
SD	117.08	114.86
Length of Association (in years)		
>0–4.5	12.1%	10.7%
>4.5–8+	87.9%	89.3%
Mean	6.72	7.01
SD	1.34	1.00

McPherson, M., Smith-Lovin, L., & Brashears, M. E. (2006). Social isolation in America: Changes in core discussion networks over two decades. American sociological review, 71(3), 353-375.

List of Other Basic Metrics

Node Degree (in-degree; out-degree)
Degree distribution
Closeness centrality
Betweenness centrality
Eigenvector centrality
Page Rank (Google)
Constraint (Structure hole)
Hubs and Authorities (HITS)
Clustering coefficient
Components
Subgraphs

N-cliques
N-clans
K-plexes
K-cores
Structural Equivalence
Shortcut
Quarter-Power Scaling
…



For more information, refer to textbooks, Wikipedia or python/R packages (e.g. NetworkX https://networkx.org/)

Newman, M. (2018). *Networks*. Oxford university press.

Advanced analysis tasks: ERGM

- If we want to study a question like, do people share misinformation more with friends than family?
  - We define a graph where nodes are people and edges indicates misinformation sharing
  - E.g., predict will there be an edge between two family members? Does homophily drive the graph structure, e.g. edges between similar nodes?
  - We could build a prediction model like logistic regression, where we try to predict an edge given the rest of the network structure, but network features are not independent!
- One solution: Exponential Random Graph Model (ERGM)
  - Assigns $P(Y = y) \propto \exp\left(\theta^{\top} s(y)\right)$,where $s(y)$ are chosen network statistics (edges, triangles, homophily, etc.) and Y is the graph we observe
  - θ are parameters we learn, e.g. through MCMC
  - We can interpret θ directly or use it for simulations

Advanced analysis tasks: Graph Neural Networks

- Problem: ERGM is inefficient for large dense graphs
- Solutions:
  - 1: **Network sampling**.
  - 2: **Transform** graph information to other data structures (e.g., vector - node embedding).
  - 3: Analyzing the graph at the local neural level and then aggregating the neurons together (e.g., **Graph Neural Network (GNN)**).
    - We can use GNNs for graph-level, node-level, or edge-level prediction tasks

Recap

- Drawing from causal inference to improve NLP models
  - Applying ideas from causal inference to model development

- Network analysis:
  - Motivation: why do we need network analysis methods?
  - Ways to represent networks
  - Example analysis: basic metrics, ERGM, GNN