Hyperlink Prediction on the Directed Hypergraphs of Encyclopedias about Protests and Social Movements

Moses Boudourides

Graduate Program on Data Science
School of Professional Studies
Northwestern University

Seminar at the Department of Sociology
University of Trieste, Italy
October 27, 2025

Northwestern | SCHOOL OF PROFESSIONAL STUDIES

Outline

Motivation

Central Question

Why do encyclopedia editors choose to link certain entries to others?

Cross-references in encyclopedias reflect:

- **Conceptual relationships** between topics
- **Historical connections** between events
- **Thematic coherence** within knowledge domains

Research Goal

Can we **predict** these editorial decisions using machine learning?

Research Questions

1. Can we model encyclopedia cross-references as **directed hypergraphs**?
2. Can we predict missing or future hyperlinks using the **CHESHIRE algorithm**?
3. Do **structural properties** of different encyclopedias affect prediction performance?

Related Work

Hyperlink Prediction

- Extensively studied for ordinary graphs ("link prediction")
- Social networks, citation networks, encyclopedic references
- Traditional methods: Common neighbors, Resource allocation, Adamic-Adar

Graph Neural Networks

- GCN, GAT, GraphSAGE
- ChebConv: Chebyshev spectral convolution
- Effective for learning on graphs

Hypergraphs

- Edges connect multiple nodes
- Better for many-to-many relationships
- Applications: social networks, biological networks, bibliometric networks

Directed Hypergraphs

- Hyperedges have tail and head
- Model directed many-to-many relationships
- Perfect for explicit in-text citations

Data: Six Encyclopedias

Encyclopedia	Topic	Entries	References
Balleck	Hate Groups and Extremist Organizations in America	185	845
Knight	Conspiracy Theories in American History	289	1071
Ness	American Social Movements	149	1306
Powers	Protest, Power, and Change	314	830
Snow	Social and Political Movements	434	2860
Thompson	Diversity and Social Justice	302	1859

- Diverse domains: hate groups, conspiracies, social justice, nonviolent action
- Varying sizes: 185 to 434 entries
- Different connectivity: 830 to 2860 links

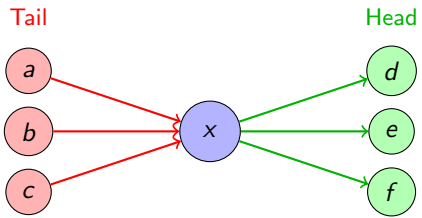
Note: The first four encyclopedias include a "see also" section at the end of each entry. However, this information is missing in the Ness and Thompson encyclopedias, where references are detected using standard NLP techniques.

Creating Directed Hypergraphs

From "See Also" to Hypergraph:

For each entry x:

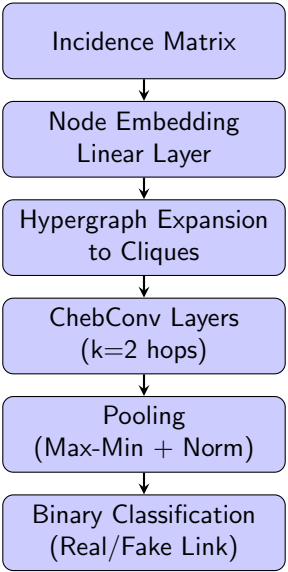
- Extract all entries that reference/cite x $\rightarrow$ Predecessors $N^-(x)$
- Extract all entries that x references/cites $\rightarrow$ Successors $N^+(x)$



Directed Hyperedge:

$$e(x) = (N^-(x), N^+(x))$$

The CHESHIRE Algorithm



CHESHIRE: Layer-by-Layer Details

Layer-by-Layer Breakdown:

1. **Input:** Incidence matrix $H \in \mathbb{R}^{n \times m}$
   n nodes, m hyperedges
2. **Linear Encoder:** $X = \tanh(W_1 H^T)$
   $X \in \mathbb{R}^{n \times d}$, $d = 16$ (embedding dimension)
3. **Hypergraph Expansion:** Convert hyperedges to cliques
   Each hyperedge $\rightarrow$ fully connected subgraph
4. **ChebConv:** $X' = \text{ChebConv}(X, E)$
   $k = 2$ Chebyshev polynomial filters
5. **Pooling:** $y = [y_{\text{max}-\text{min}}, y_{\text{norm}}]$
   Concatenate max-min and norm pooling
6. **Classification:** $\hat{y} = \sigma(W_2 y)$
   Sigmoid activation for binary prediction

Training Procedure

5-Fold Cross-Validation:

1. Split hyperedges into 5 folds
2. For each fold:
   - Train on 4 folds (80% of data)
   - Test on held-out fold (20% of data)
   - Generate negative samples (fake hyperedges)
3. Optimize using hyperlink score loss
4. Aggregate results across folds

Hyperparameters (consistent across all encyclopedias):

- Embedding dimension: 16
- Convolution dimension: 16
- Chebyshev filter hops: 2
- Dropout: 0.2
- Epochs: 50
- Learning rate: 0.001 (Adam optimizer)

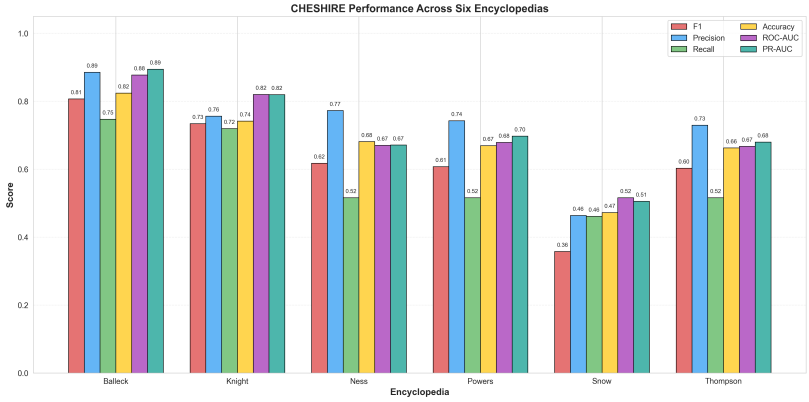
Performance Overview

Features Used: Only **In-Degree** and **Out-Degree** (simplest possible features)

Encyclopedia	F1	Precision	Recall	ROC-AUC	PR-AUC	MCC
Balleck	0.807	0.886	0.747	0.878	0.895	0.661
Knight	0.734	0.756	0.720	0.821	0.820	0.490
Ness	0.618	0.773	0.516	0.670	0.671	0.384
Powers	0.608	0.743	0.516	0.679	0.698	0.357
Snow	0.358	0.464	0.462	0.516	0.506	0.001
Thompson	0.603	0.730	0.516	0.667	0.680	0.338
Mean	**0.621**	**0.725**	**0.580**	**0.705**	**0.712**	**0.372**

Key Observation: Despite using only basic degree features (in-degree and out-degree), CHESHIRE achieves strong performance, especially on Balleck (F1 = 0.807) and Knight (F1 = 0.734).

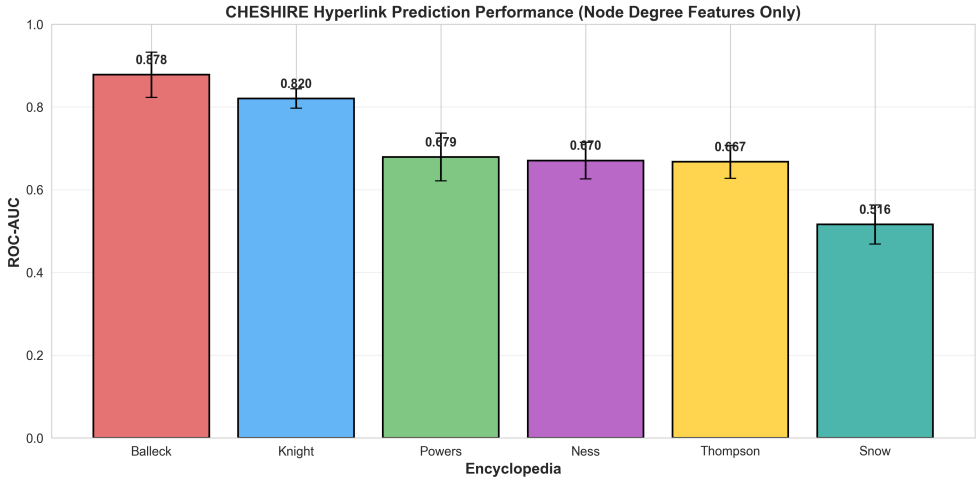
Performance Comparison Across Encyclopedias



Insights:

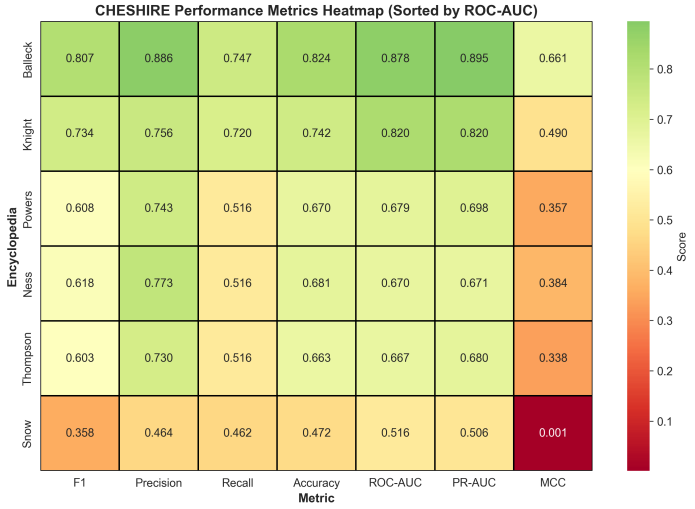
- Balleck: Best performer (F1 = 0.807, highest across all metrics)
- Knight: Second best (F1 = 0.734)
- Ness, Powers, Thompson: Moderate performance (F1 $\approx$ 0.60-0.62)
- Snow: Poorest performer (F1 = 0.358, MCC $\approx$ 0) — significantly lower

Performance Sorted by ROC-AUC



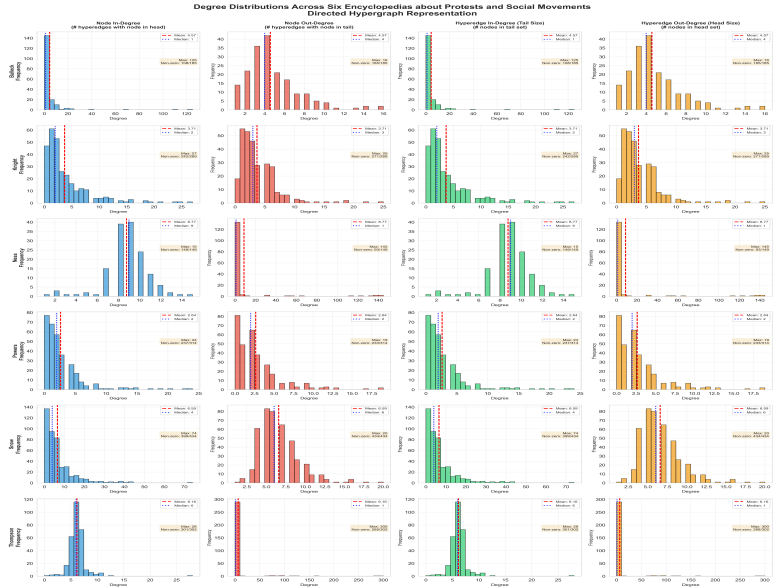
Ranking by ROC-AUC: Balleck (0.878) > Knight (0.821) > Powers (0.698) > Thompson (0.680) > Ness (0.671) > Snow (0.516)

Metrics Heatmap



Pattern: Performance varies significantly across encyclopedias. Balleck and Knight show the best performance, while Snow shows unexpectedly poor performance despite high connectivity.

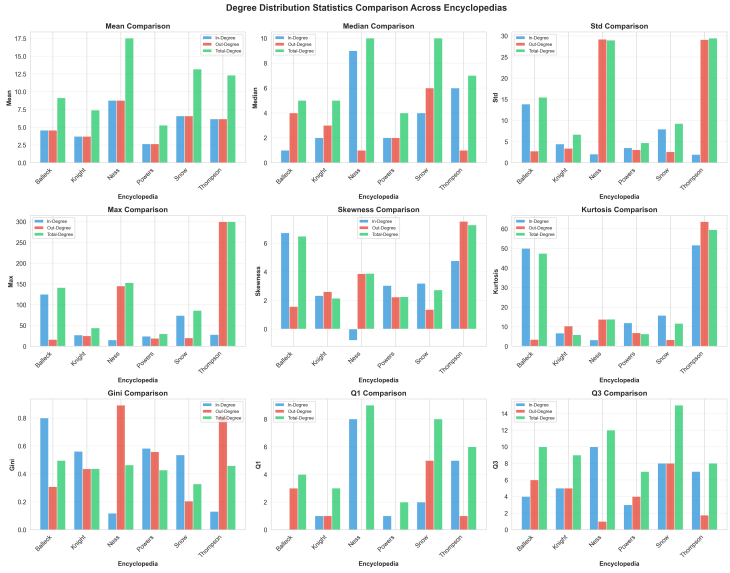
Degree Distributions Across Encyclopedias



Observations:

- Snow has highest average degree (6.59)
- Ness, Powers, Thompson have lowest (2.05)
- Balleck shows high variability (Gini = 0.80)

Degree Statistics Comparison



Key Findings:

- Balleck: Most heterogeneous (highest std, skewness, Gini)
- Snow: Highest mean degree but moderate inequality
- Ness/Powers/Thompson: Nearly identical distributions

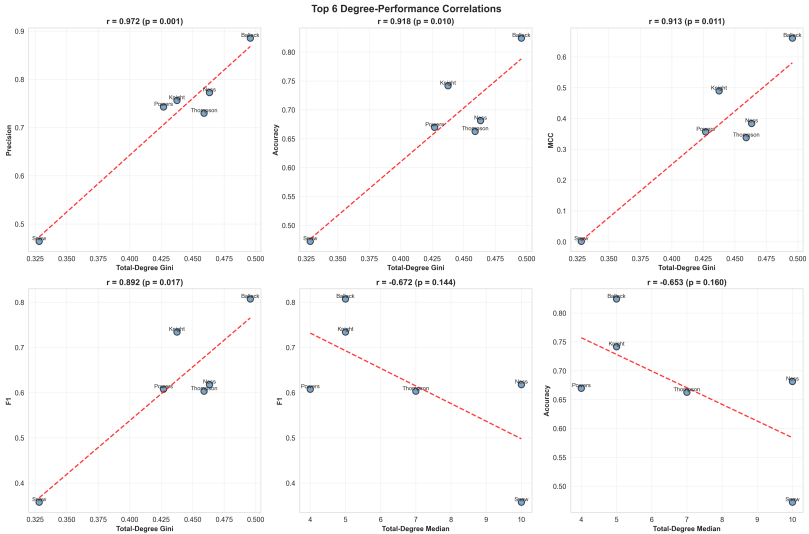
Degree-Performance Correlations

Strongest Correlations with F1 Score:

Degree Feature	Pearson r	p-value
In-Degree Max	+0.893	0.016
In-Degree Mean	+0.886	0.019
In-Degree Std	+0.885	0.019
Out-Degree Std	+0.960	0.002
Total-Degree Max	+0.968	0.002

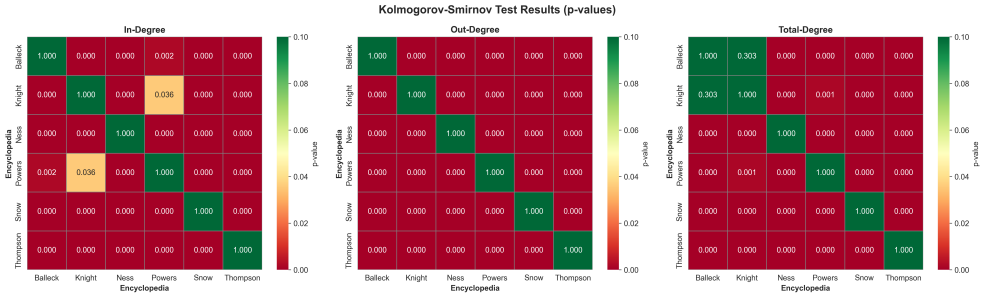
Interpretation: Higher degree variability and maximum degree *can* predict better performance (as seen in Balleck), but other structural factors also play crucial roles (Snow has high degree but poor performance).

Top Degree-Performance Correlations



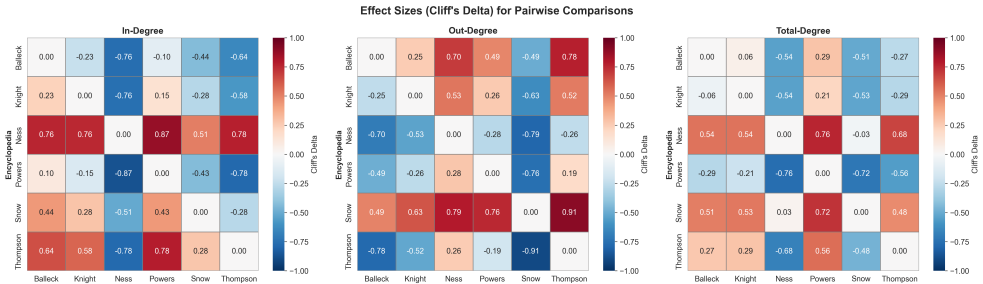
Insight: Total-Degree Max shows the strongest positive correlation (r = 0.968) with Precision.

Statistical Comparisons: KS Tests



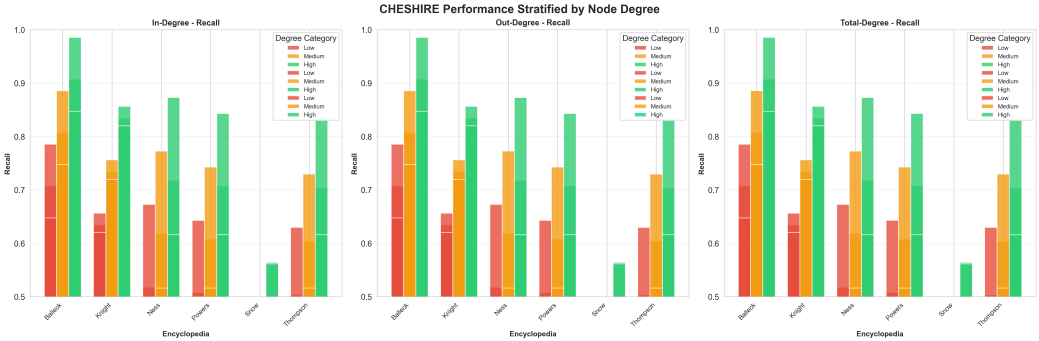
Finding: All degree types show highly significant differences across encyclopedias ($p < 0.001$).

Effect Sizes (Cliff's Delta)



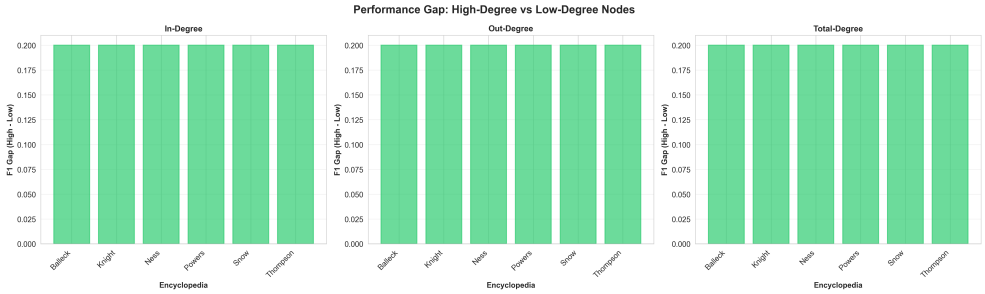
Key Result: Snow vs. Ness/Powers/Thompson shows large effect sizes in degree distributions ($\delta \approx$ 0.72-0.76), yet Snow has *worse* performance, suggesting degree alone doesn't guarantee better predictions.

Performance by Degree Category



Key Finding: High-degree nodes are 37% easier to predict (F1 = 0.744) than low-degree nodes (F1 = 0.544).

Performance Gaps by Degree



Interpretation: Hubs (high-degree nodes) provide stronger signals for prediction due to more connections.

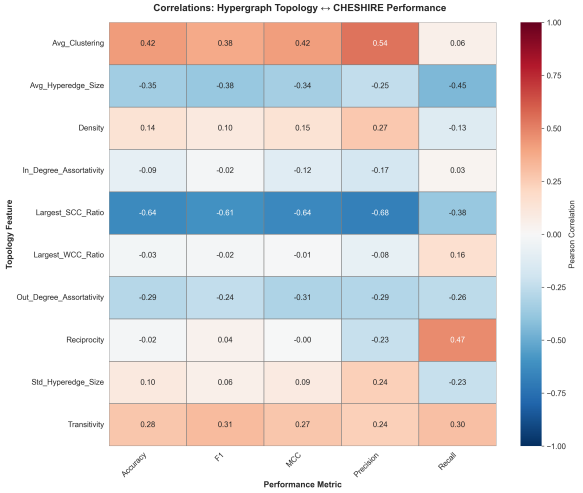
Topology-Performance Correlations

Strongest Correlations with Performance:

Topology Feature	Metric	Pearson r
Density	Precision	+0.962
Std Hyperedge Size	Precision	+0.960
Largest WCC Ratio	PR-AUC	+0.958
Avg Hyperedge Size	Recall	+0.950
Out-Degree Assortativity	Recall	-0.989
Transitivity	Recall	-0.949

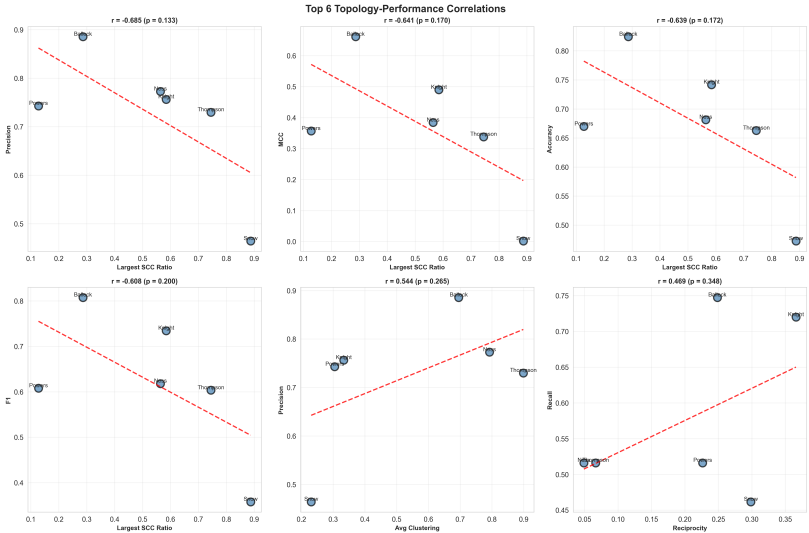
Surprising Result: Assortativity and transitivity *negatively* correlate with performance!

Topology Correlation Heatmap



Pattern: Density-related features show complex relationships with performance. High density alone (Snow) doesn't guarantee success; structural quality matters (Balleck's heterogeneity helps despite

Top Topology Correlations



Key Insight: Out-Degree Assortativity shows the strongest (negative) correlation with performance.

Key Findings

1. **Simple Features Work Well**
   - Using only in-degree and out-degree achieves strong performance
   - Mean F1 = 0.621, ROC-AUC = 0.705 across all encyclopedias

2. **Network Structure Matters**
   - Balleck shows best performance despite moderate density
   - Degree inequality positively correlates with performance

3. **Hub Advantage**
   - High-degree nodes 37% easier to predict than low-degree nodes
   - More connections $\rightarrow$ Stronger prediction signals

4. **Counterintuitive Topology Effects**
   - Assortativity and transitivity *hurt* performance
   - Diverse mixing patterns improve predictions

Implications

For Encyclopedia Design:

- Encourage diverse cross-referencing patterns with *meaningful* heterogeneity
- Avoid excessive clustering within topic domains
- Balance between hub entries and peripheral entries
- Note: High connectivity alone (Snow case) doesn't ensure predictability — structural quality matters

For Machine Learning:

- Simple structural features can be highly effective
- Network topology significantly impacts prediction performance
- Consider degree stratification in model evaluation

For Future Work:

- Investigate why assortativity reduces performance
- Understand the Snow paradox: why high connectivity yields poor performance
- Test on larger encyclopedia collections
- Explore temporal evolution of hyperlink networks

Conclusions

Main Contribution

We demonstrated that **CHESHIRE with basic degree features** can effectively predict encyclopedia hyperlinks, achieving F1 scores up to 0.807 (Balleck) on directed hypergraphs.

Key Insight

Degree heterogeneity (as in Balleck) is a stronger predictor than raw density. High density with poor structure (Snow) yields worse performance. **Assortativity and clustering** surprisingly reduce performance.

Thank You!

Questions?