Exploratory Analysis of Graph Data by Leveraging Domain Knowledge

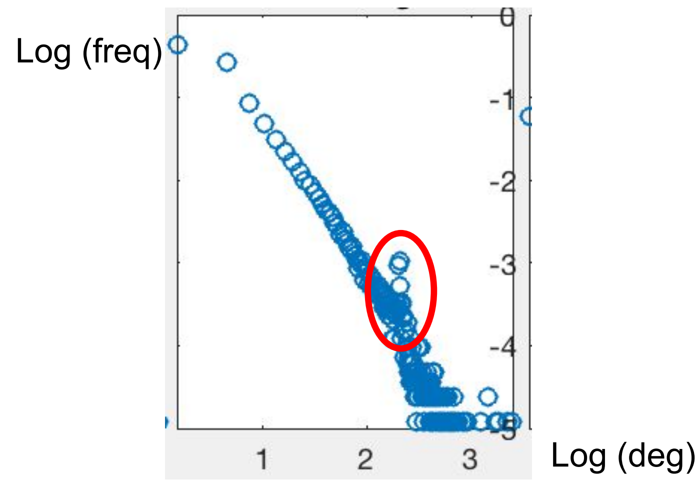
Di Jin

Danai Koutra

IEEE International Conference on Data Mining (ICDM), 2017

Graph invariants are prevalent

- In many tasks (e.g., anomaly detection, classification, ...)
  - Different invariants
    - Degrees
    - Betweenness
    - Average path length
    - Giant components
    - … ...
  - Compare them with “common” laws
    - The power-like laws
    - 6 degree of separation (log(N) / log(c))
    - 1 giant component
    - … ...



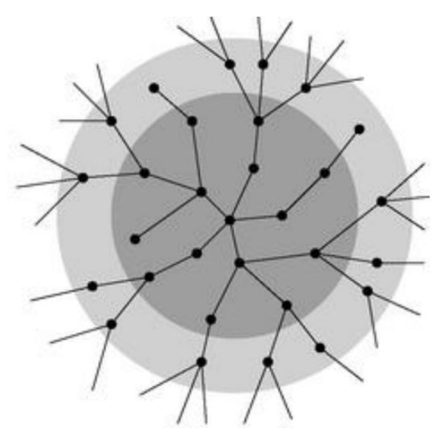
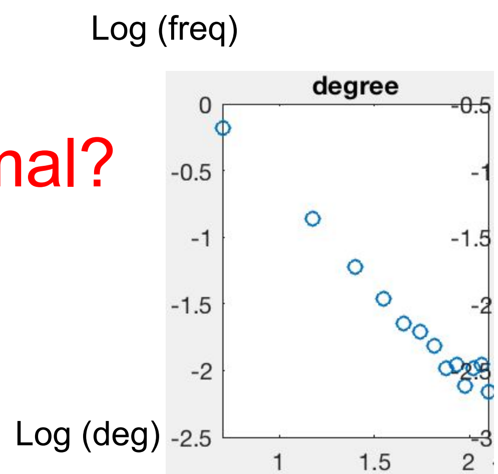


Image source: Newman, Mark. *Networks: an introduction*.

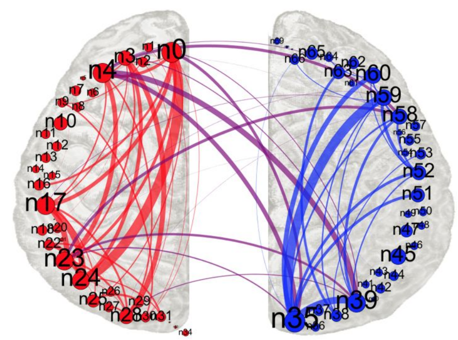
Are graph invariants enough to understand?

- The brain connectivity correlation graph

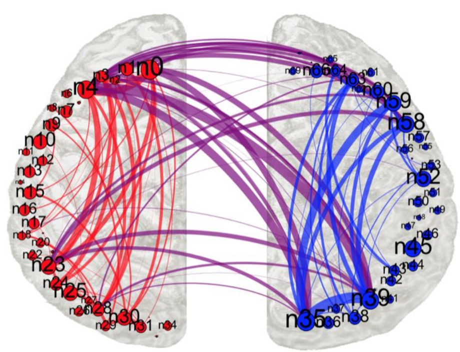
Normal?

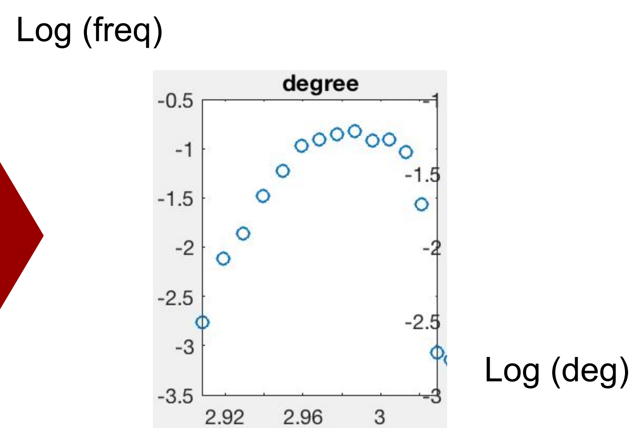


Patient?



Healthy subject





Are graph invariants enough to understand?

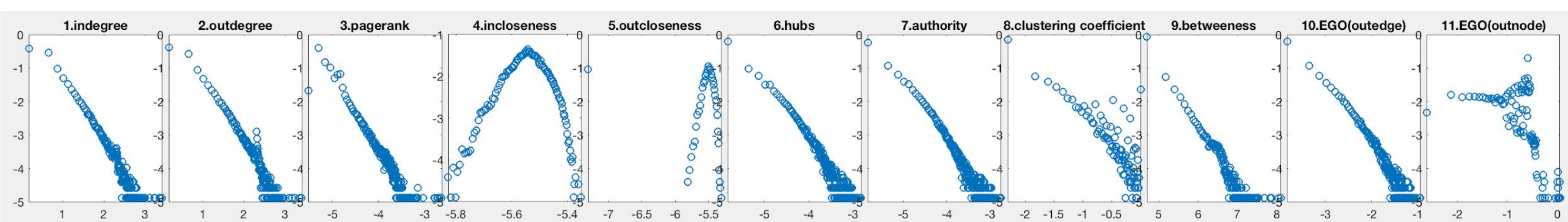
- Common laws are not golden
  - Graph invariant distributions are good.
  - But bigger picture should not be neglected.

Are graph invariants enough to understand?

- Common laws are not golden
  - Graph invariant distributions are good.
  - But bigger picture should not be neglected.
- Prior/Domain knowledge is important
  - Graphs are everywhere, but the domain experts are NOT.
  - “What patterns are expected?”

Are graph invariants enough to understand?

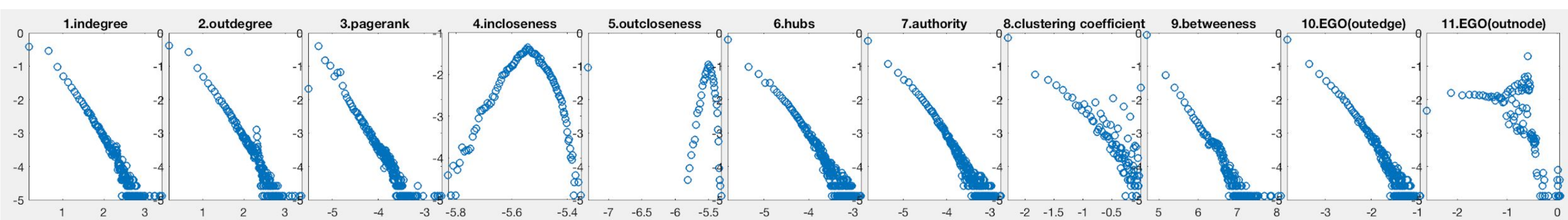
- Common laws are not golden
  - Graph invariant distributions are good.
  - But bigger picture should not be neglected.
- Prior/Domain knowledge is important
  - Graphs are everywhere, but the domain experts are NOT.
  - “What patterns are expected?”
- Useful graph invariant distributions (**features**) vary
  - Tons of features can be extracted, few are useful
  - “Which features to explore?”



…. ….

Are graph invariants enough to understand?

- Common laws are not golden
  - Graph invariant distributions are good.
  - But bigger picture should not be neglected.
- Prior/Domain knowledge is important
  - Graphs are everywhere, but the domain experts are NOT.
  - **"What patterns are expected?"**
- Useful graph invariant distributions (**features**) vary
  - Tons of features can be extracted, few are useful
  - **"Which features to explore?"**

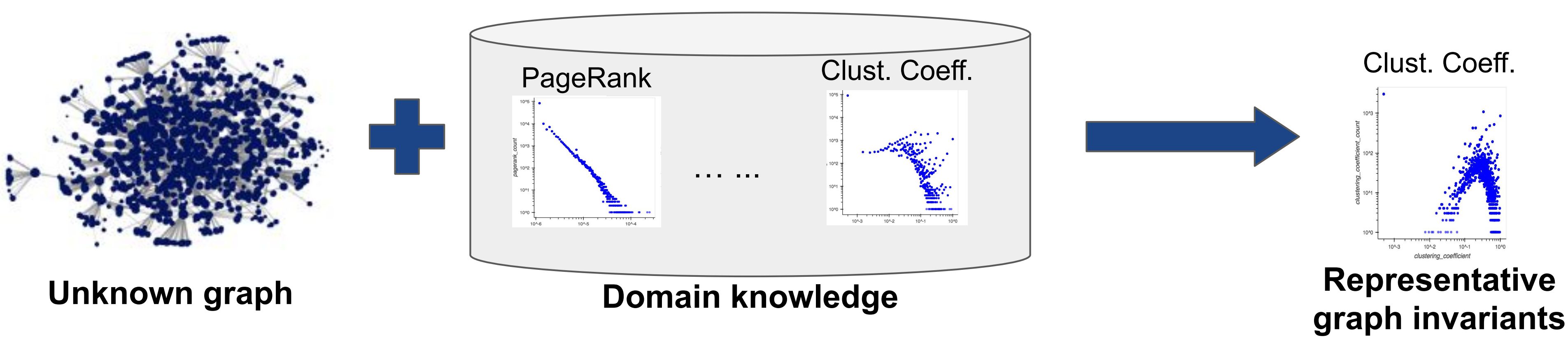


…. ….

EAGLE: Exploratory Analysis of Graphs with domain knowLEdge

Given: an input graph & *domain knowledge*
Find: brief summary consisting of **representative** features that satisfies a set of desired properties (e.g., diversity)

- **“What patterns are expected?”**
  - *Domain* knowledge: a collection of graphs with all features in the feature space.
- **“Which features to explore?”**
  - *Representative* features: graph invariant distributions (PDF) with desired properties.



Many Existing Methods

Check the centralities!

PageRank

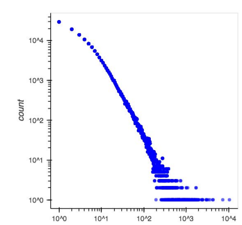
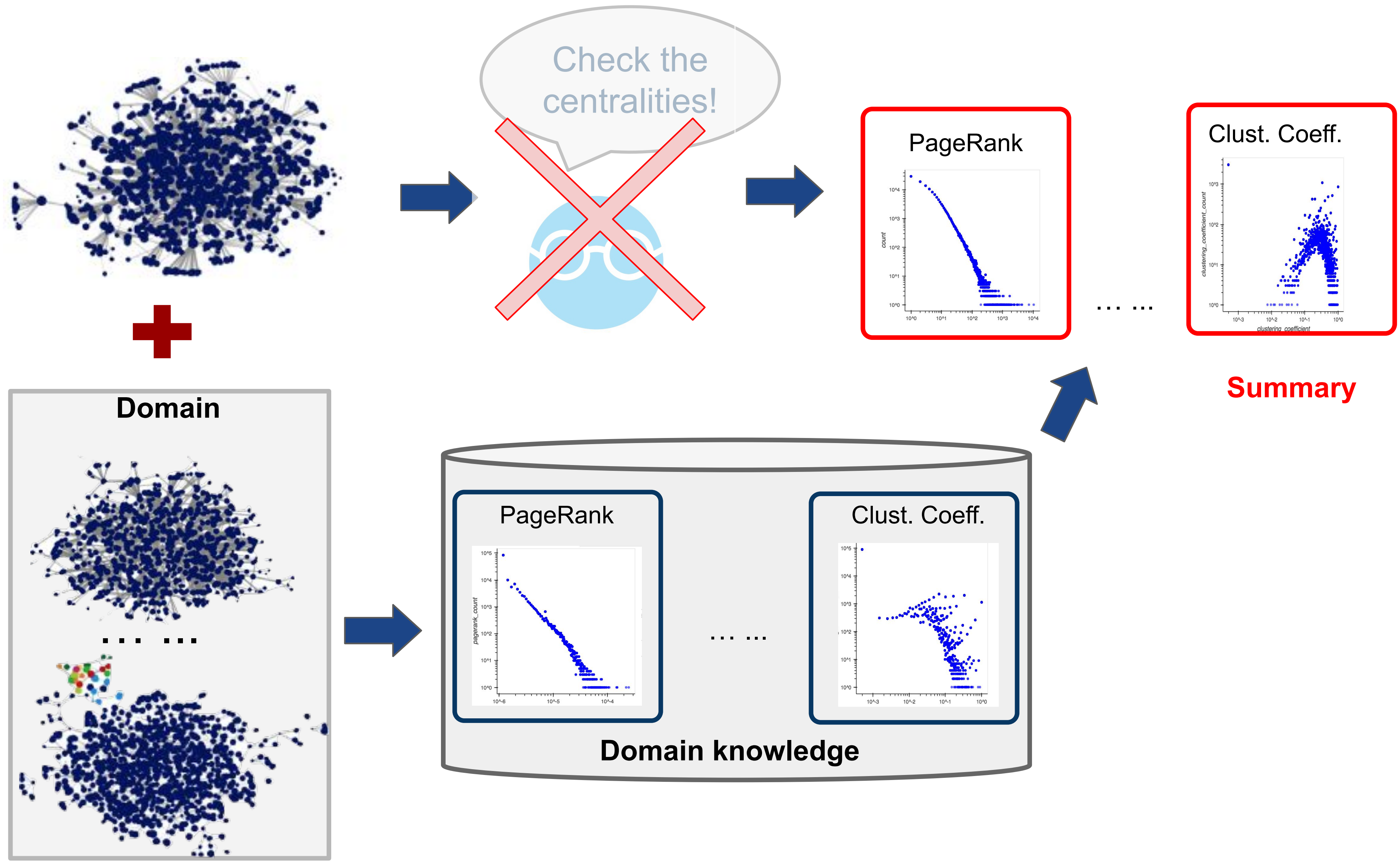


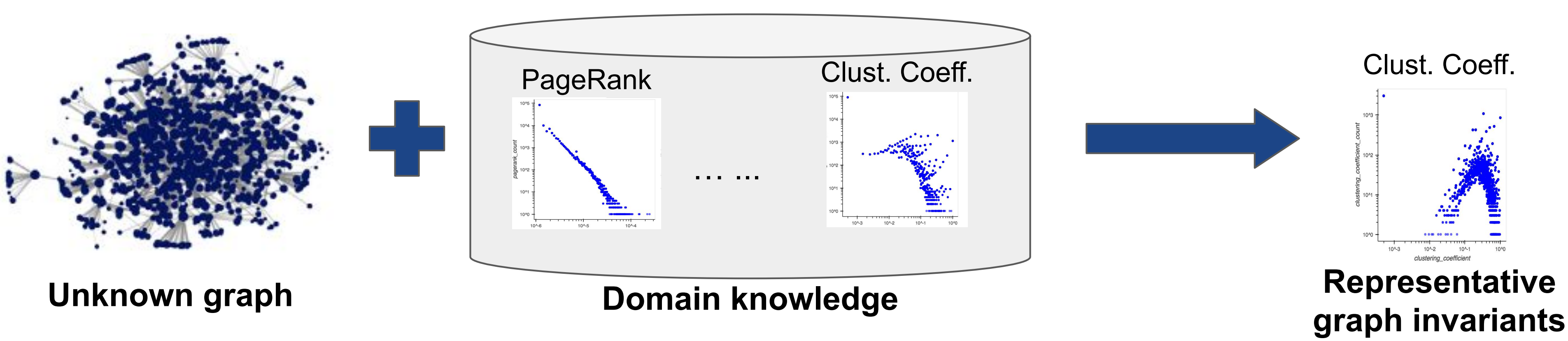
Image source: Duan, Wei, et al. "Topology dependent epidemic spreading velocity in weighted networks."

Proposed Solution: key idea



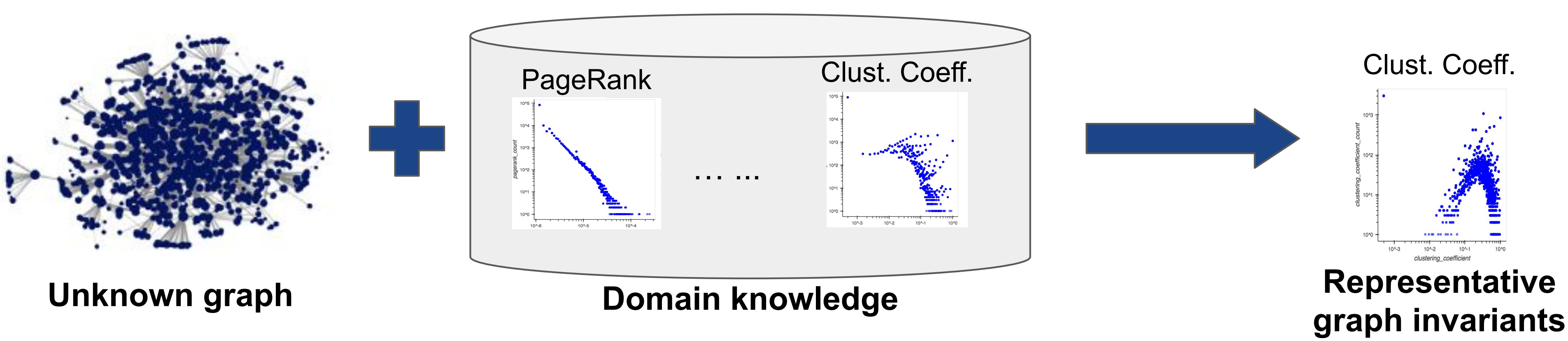
Proposed Solution: key idea

- **“Summarize an unknown graph from known ones”**.
  - “Known graphs”**:** the *domain knowledge*.
  - Summarize through *representative* graph invariants.
  - Discovers domain-specific patterns **automatically.**



Proposed Solution: key idea

- **"Summarize an unknown graph from known ones".**
  - "Known graphs"**:** the *domain knowledge*.
  - Summarize through *representative* graph invariants.
  - Discovers domain-specific patterns **automatically.**
- Not a traditional graph summarization problem.
  - No compressed representation of an input graph.



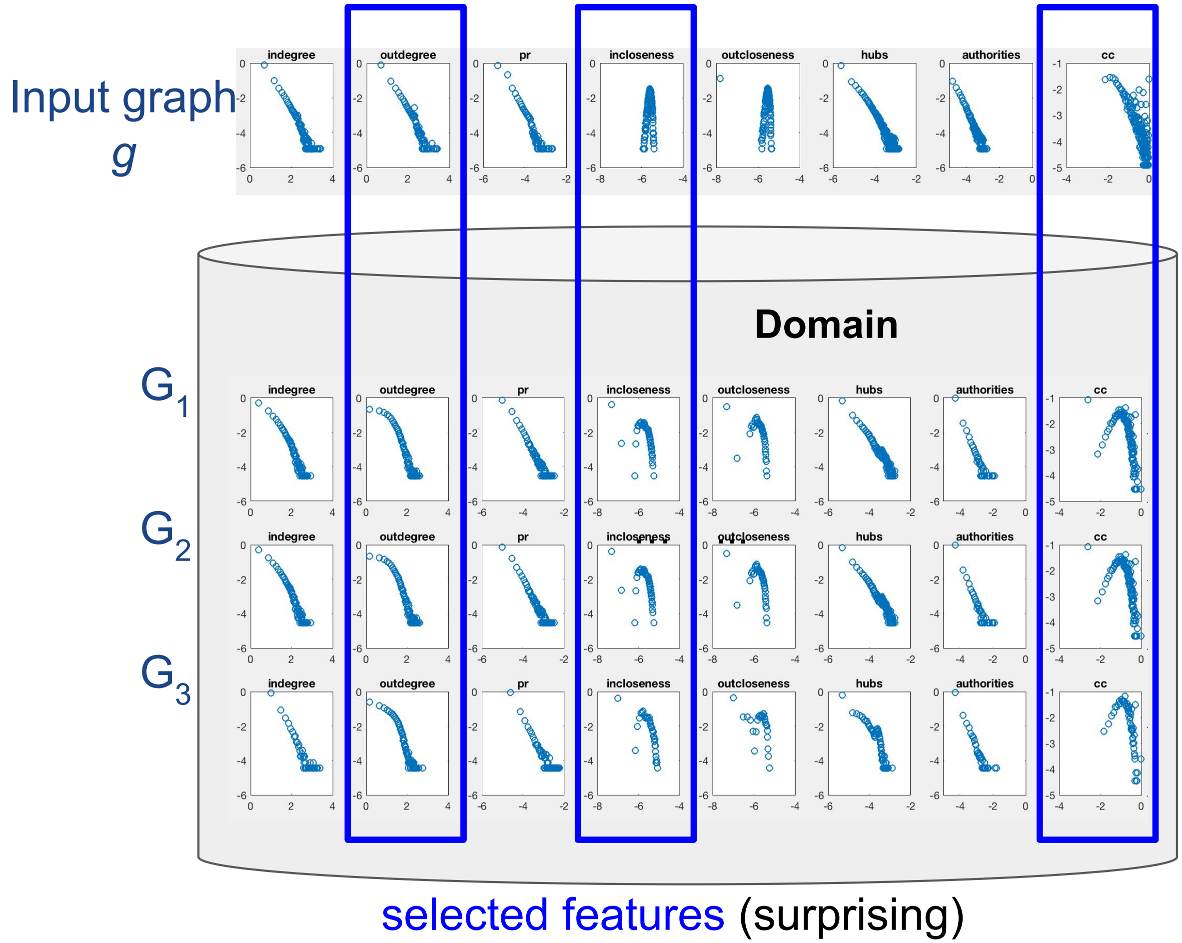
EAGLE: Desired properties

The **summary** should be:

- Diverse
- Concise
- Domain-specific

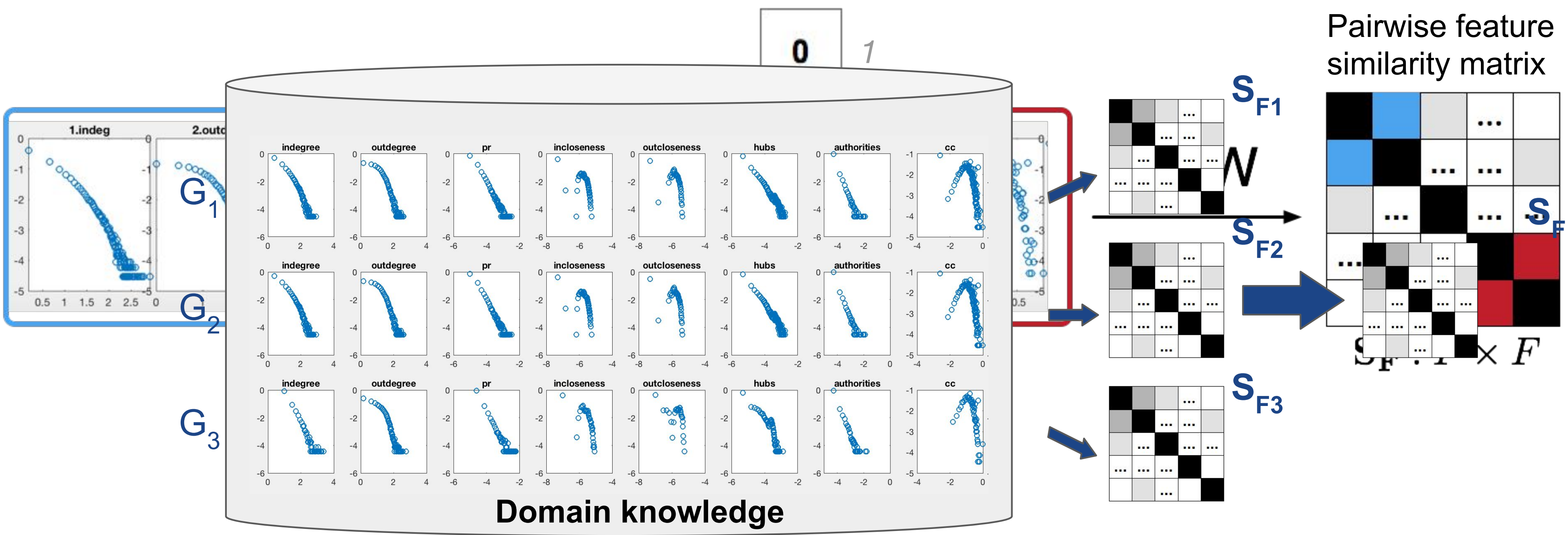
The **process** should be:

- Efficient
- Interpretable: univariate graph statistics (PDF)



EAGLE: Formulation

$$\arg\min_{\mathbf{f}} \lambda_1 \boxed{\underbrace{\mathbf{f}^T \mathbf{S_F} \mathbf{f}}_{\text{Diversity}}} + \lambda_2 \underbrace{\|\mathbf{f}\|_0}_{\text{Conciseness}} + \lambda_3 \cdot \underbrace{\mathbf{f}^T \mathbf{h}}_{\text{Domain-specificity}}$$



$$\mathbf{S_F}(f_j, f_l) = \sum_{i=1}^{K} w_i \cdot \mathbf{S_{Fi}}(f_j, f_l)$$

EAGLE: Formulation

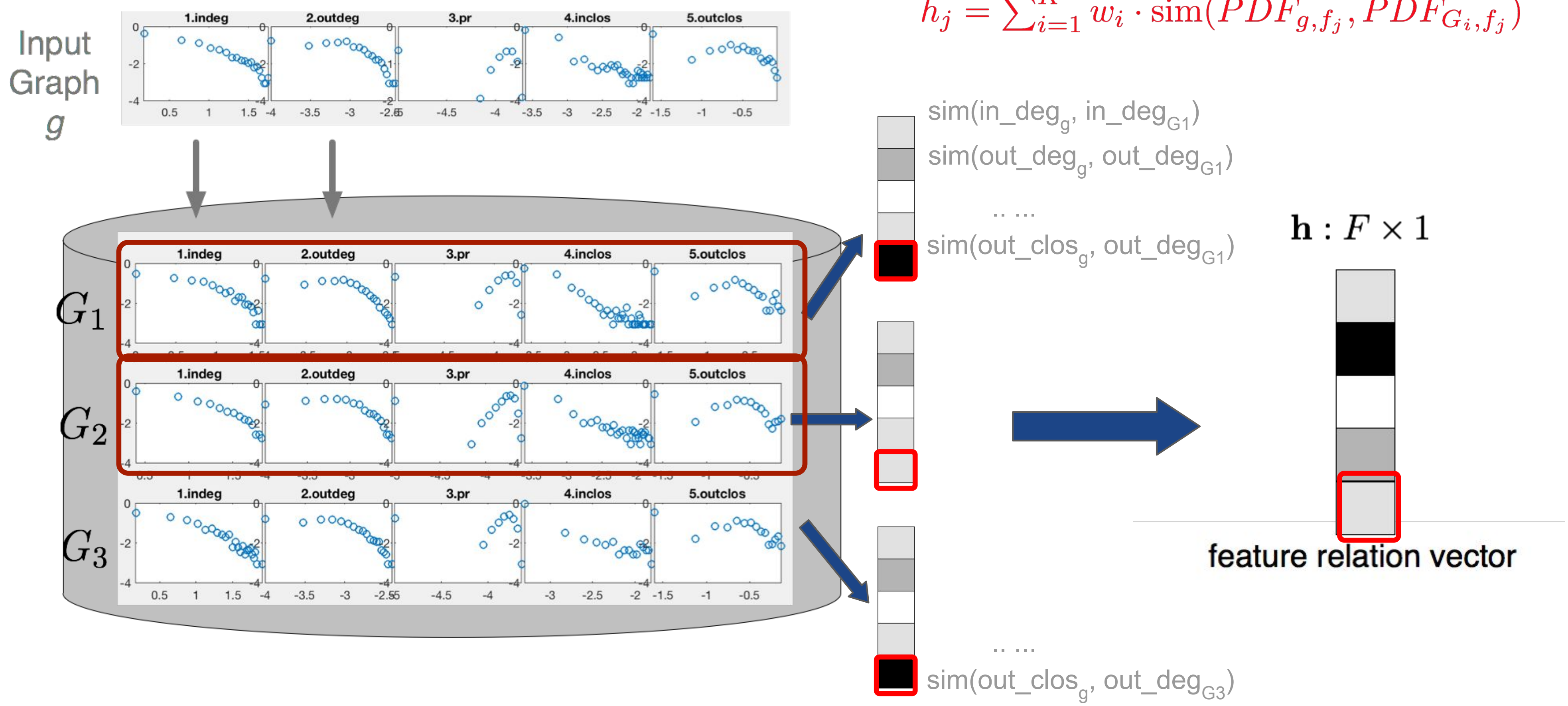
$$\arg\min_{\mathbf{f}} \lambda_1 \underbrace{\mathbf{f}^T \mathbf{S_F} \mathbf{f}}_{\text{Diversity}} + \lambda_2 \boxed{\underbrace{||\mathbf{f}||_0}_{\text{Conciseness}}} + \lambda_3 \cdot \underbrace{\mathbf{f}^T \mathbf{h}}_{\text{Domain-specificity}}$$

$$||\mathbf{f}||_0 \xrightarrow{Relax} ||\mathbf{f}||_2 = \mathbf{f}^T \mathbf{f}$$

EAGLE: Formulation

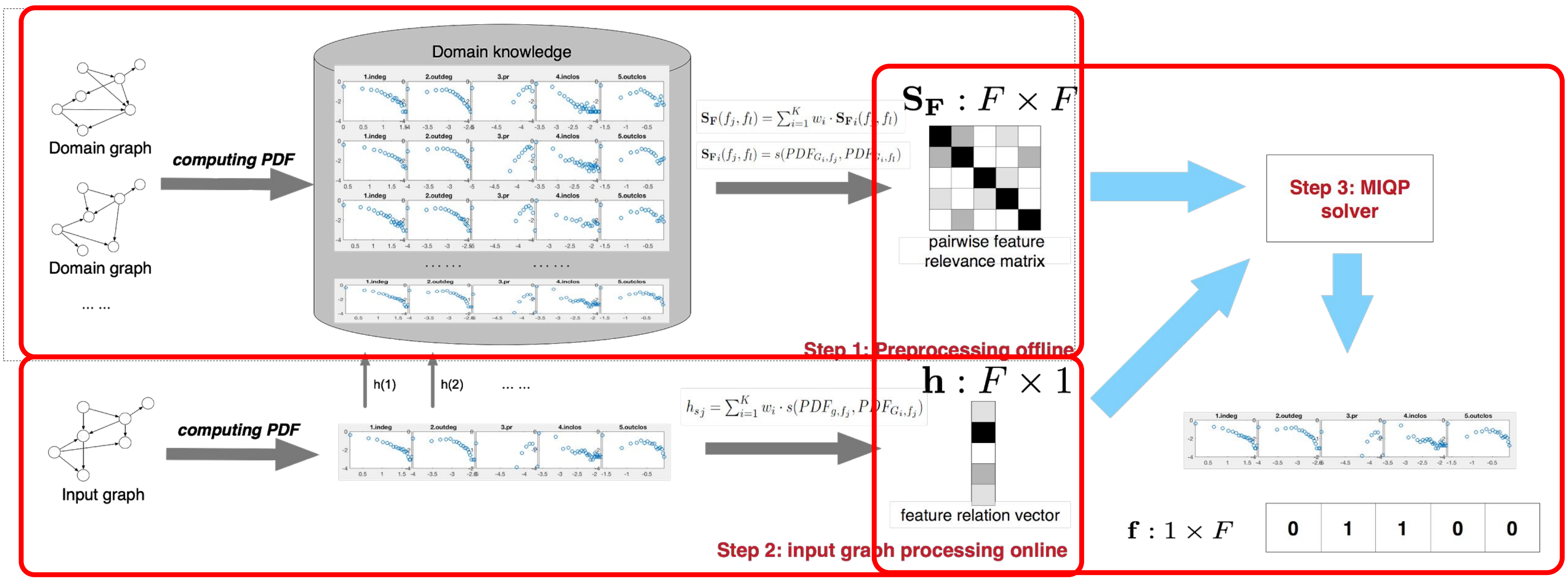
$$\arg\min_{\mathbf{f}} \lambda_1 \underbrace{\mathbf{f}^T \mathbf{S_F} \mathbf{f}}_{\text{Diversity}} + \lambda_2 \underbrace{\|\mathbf{f}\|_0}_{\text{Conciseness}} + \lambda_3 \cdot \boxed{\underbrace{\mathbf{f}^T \mathbf{h}}_{\text{Domain-specificity}}}$$

$$h_j = \sum_{i=1}^{K} w_i \cdot \mathrm{sim}(PDF_{g,f_j}, PDF_{G_i,f_j})$$



Workflow: 3 steps

$$\arg\min_{\mathbf{f}} \lambda_1 \underbrace{\mathbf{f}^T \mathbf{S_F} \mathbf{f}}_{\text{Diversity}} + \lambda_2 \underbrace{\|\mathbf{f}\|_0}_{\text{Conciseness}} + \lambda_3 \cdot \underbrace{\mathbf{f}^T \mathbf{h}}_{\text{Domain-specificity}}$$



How to solve? MIQP solver

- Before

$$\arg\min_{\mathbf{f}} \lambda_1 \underbrace{\mathbf{f}^T \mathbf{S}_\mathbf{F} \mathbf{f}}_{\text{Diversity}} + \lambda_2 \underbrace{\|\mathbf{f}\|_0}_{\text{Conciseness}} + \lambda_3 \cdot \underbrace{\mathbf{f}^T \mathbf{h}}_{\text{Domain-specificity}}$$

- Rewriting relaxed form

$$\begin{aligned} &\arg \min_{\mathbf{f} \in \{0,1\}^{F \times 1}} \lambda_1 \mathbf{f}^T \mathbf{S}_\mathbf{F} \mathbf{f} + \lambda_2 \mathbf{f}^T \mathbf{I}_F \mathbf{f} + \lambda_3 \mathbf{f}^T \mathbf{h} \\ &= \min_{\mathbf{f} \in \{0,1\}^{F \times 1}} \mathbf{f}^T \underbrace{(\lambda_1 \mathbf{S}_\mathbf{F} + \lambda_2 \mathbf{I}_F)}_{\mathbf{Q}} \mathbf{f} + \mathbf{f}^T \underbrace{\lambda_3 \mathbf{h}}_{\mathbf{r}} \end{aligned}$$

How to solve? MIQP solver

- Before

$$\arg\min_{\mathbf{f}} \lambda_1 \underbrace{\mathbf{f}^T \mathbf{S_F} \mathbf{f}}_{\text{Diversity}} + \lambda_2 \underbrace{\|\mathbf{f}\|_0}_{\text{Conciseness}} + \lambda_3 \cdot \underbrace{\mathbf{f}^T \mathbf{h}}_{\text{Domain-specificity}}$$

- Rewriting relaxed form

$$\arg \min_{\mathbf{f}\in\{0,1\}^{F\times 1}} \lambda_1 \mathbf{f}^T \mathbf{S_F} \mathbf{f} + \lambda_2 \mathbf{f}^T \mathbf{I}_F \mathbf{f} + \lambda_3 \mathbf{f}^T \mathbf{h}$$

$$= \min_{\mathbf{f}\in\{0,1\}^{F\times 1}} \mathbf{f}^T \underbrace{(\lambda_1 \mathbf{S_F} + \lambda_2 \mathbf{I}_F)}_{\mathbf{Q}} \mathbf{f} + \mathbf{f}^T \underbrace{\lambda_3 \mathbf{h}}_{\mathbf{r}}$$

- The general form: MIQP & MILP with slack variable z

$$\begin{aligned} \underset{\mathbf{f}}{\text{minimize}} \quad & \mathbf{f}^T \mathbf{Q} \mathbf{f} + \mathbf{r}^T \mathbf{f} \\ \text{subject to} \quad & 0 \leq \sum_i^F \mathbf{f}(i) \leq F \\ & 0 \leq \mathbf{f}(i) \leq 1, \; i = 1, \ldots, F. \end{aligned}$$

Slack variable z

$$\begin{aligned} \underset{\mathbf{f},z}{\text{minimize}} \quad & z + \mathbf{r}^T \mathbf{f} \\ \text{subject to} \quad & 0 \leq \sum_i^F \mathbf{f}(i) \leq F \\ & 0 \leq \mathbf{f}(i) \leq 1, \; i = 1, \ldots, F. \\ & \mathbf{f}^T \mathbf{Q} \mathbf{f} - z \leq 0, \; z \geq 0 \end{aligned}$$

"0-pit" Problem

$$\arg \min_{\mathbf{f} \in \{0,1\}^{F \times 1}} \lambda_1 \mathbf{f}^T \mathbf{S_F} \mathbf{f} + \lambda_2 \mathbf{f}^T \mathbf{I}_F \mathbf{f} + \lambda_3 \mathbf{f}^T \mathbf{h}$$

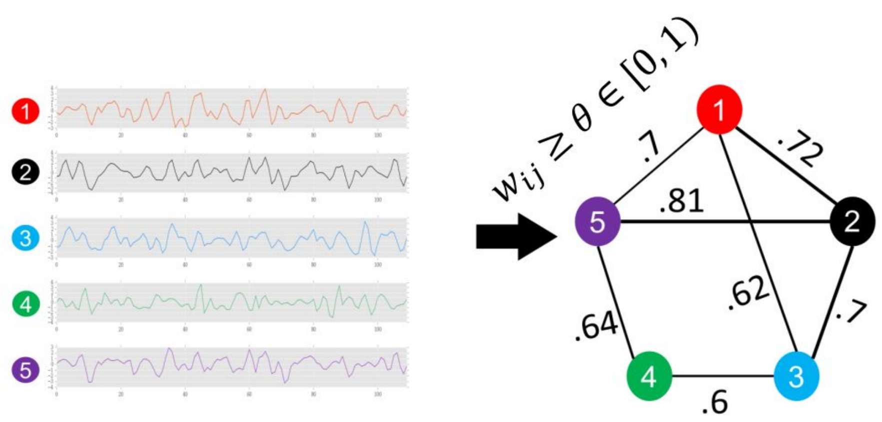
$$= \min_{\mathbf{f} \in \{0,1\}^{F \times 1}} \mathbf{f}^T \underbrace{(\boxed{\lambda_1}\mathbf{S_F} + \boxed{\lambda_2}\mathbf{I}_F)}_{\mathbf{Q}} \mathbf{f} + \mathbf{f}^T \underbrace{\boxed{\lambda_3}\mathbf{h}}_{\mathbf{r}}$$

- "0-pit" problem:
  - All the terms are positive
  - Optimal solution: for **f** = all-**0** vector
- Solution:
  - EAGLE-Fix: Explicitly set the # of selected features in **f**
  - EAGLE-Flex: Set negative value to the normalization term

Experiments: data

- Feature space (28 in total):
  - **Node-specific**: in/out-degree, PageRank, hubs, authorities, roles, …
  - **Structure-specific**: # in/out neighbors and # in/out edges of egonets, …, distribution of communities, motifs, …

Domain	Name	Nodes	Edges	Description
Connectomics	Brain-Voxel1	3 789	399 069	directed unweighted
	Brain-Voxel2	3 789	148 648	directed unweighted
Citation networks	HepTh	27 770	352 807	directed unweighted
	HepPh	34 546	421 578	directed unweighted
Social science	Epinions	75 879	508 837	directed unweighted
	Slashdot0811	77 360	905 468	directed unweighted
	Slashdot0922	82 168	948 464	directed unweighted



Experiments: baseline methods

No existing domain-specific summarization method!

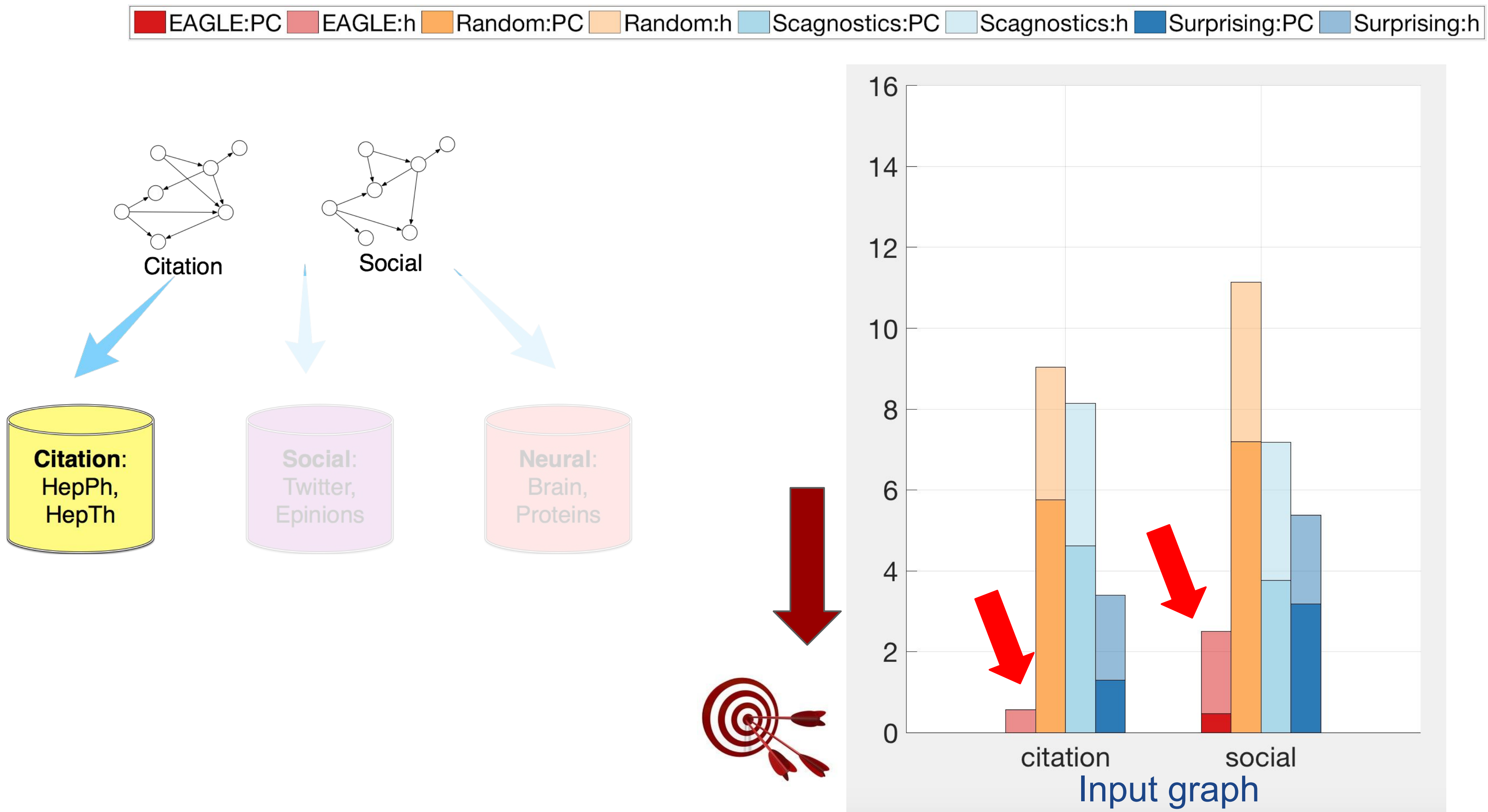
- Random selection
- Surprising selection

$$\arg\min_{\mathbf{f}} \lambda_1 \underbrace{\mathbf{f}^T \mathbf{S_F} \mathbf{f}}_{\text{Diversity}} + \lambda_2 \underbrace{\|\mathbf{f}\|_0}_{\text{Conciseness}} + \lambda_3 \cdot \underbrace{\mathbf{f}^T \mathbf{h}}_{\text{Domain-specificity}}$$

- SCAGNOSTICS
  - Pick each feature *independently* based on its anomalies in density, shape and trend.
  - 9 scores: stringiness, skewness, skinniness, etc.

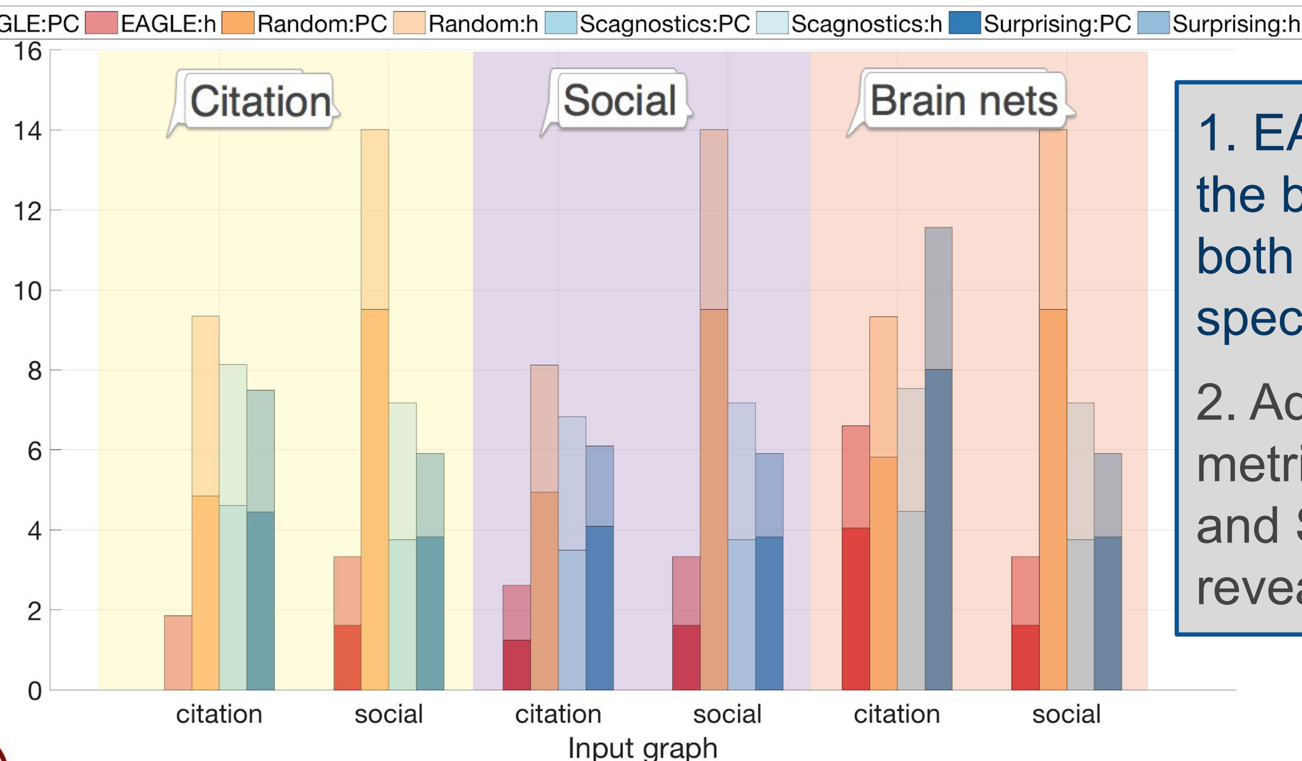
Diversity & Domain-specificity

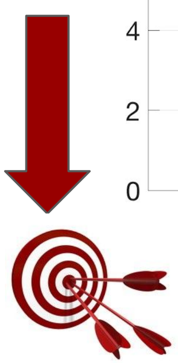
- **Metric:** Pearson correlation (PC)



Diversity & Domain-specificity

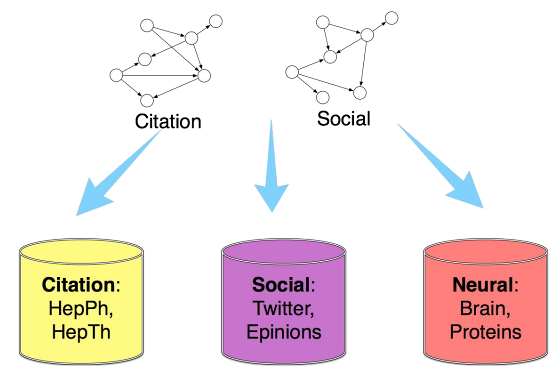
- **Metric:** Pearson correlation (PC)



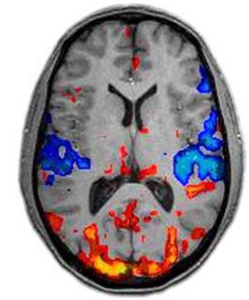


1. EAGLE outperforms the baselines in terms of both diversity and domain specificity.

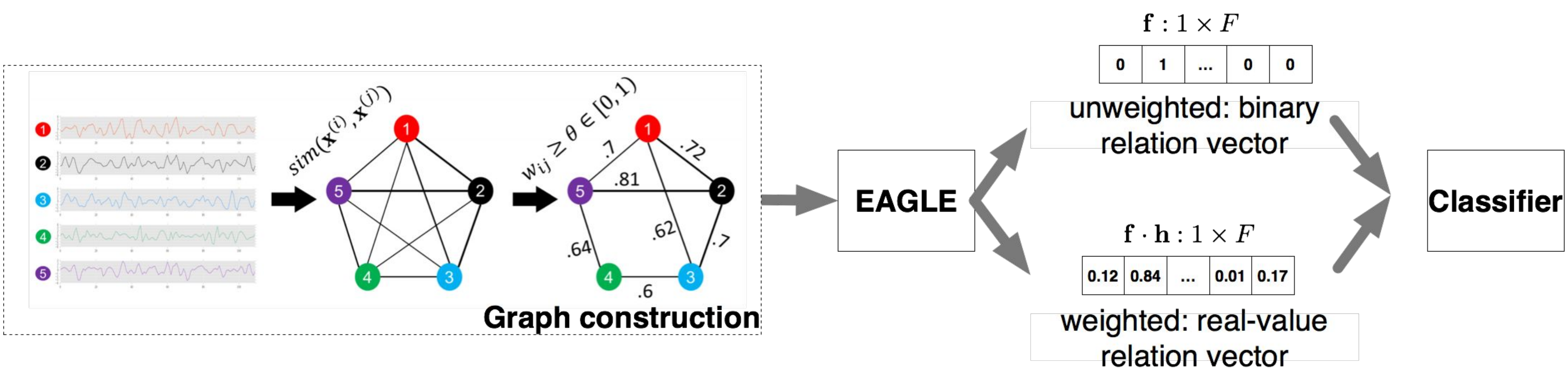
2. Additional evaluation metrics (Kendall's Tau and Spearman's Rank) reveal similar patterns.



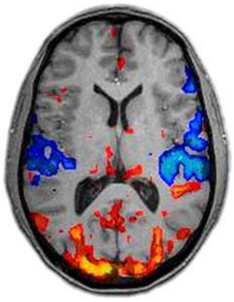
Case study: brain graph classification



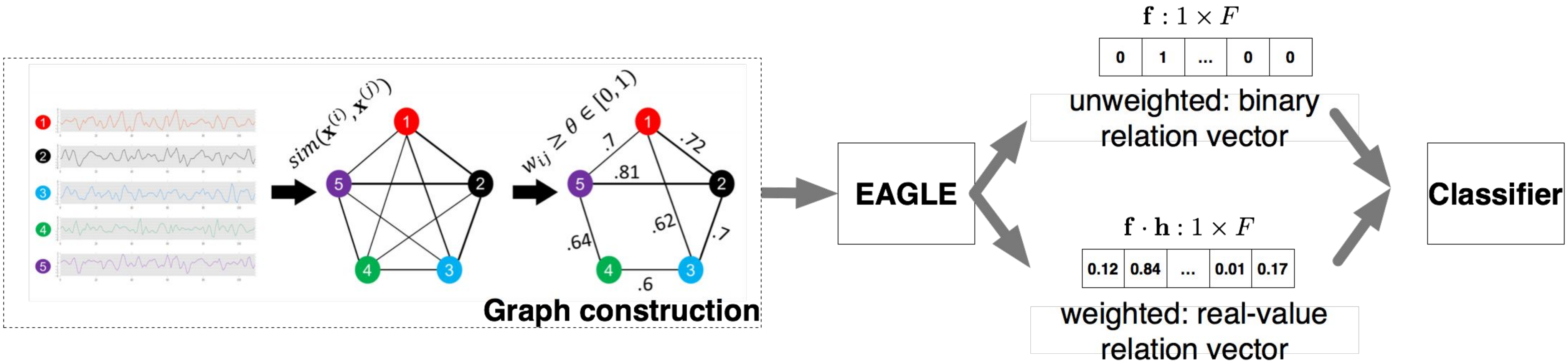
- Setup (EAGLE-Fix & -Flex)
  - COBRE dataset: 72 patients with schizophrenia and 76 healthy controls, 1166 fMRI time series.
  - Threshold: 0.6
  - Feature space: 11 (degree, clustering coeff, betweenness, …)



Case study: brain graph classification



- Setup (EAGLE-Fix & -Flex)
  - COBRE dataset: 72 patients with schizophrenia and 76 healthy controls, 1166 fMRI time series.
  - Threshold: 0.6
  - Feature space: 11 (degree, clustering coeff, betweenness, …)
- Baselines
  - Baseline 1: average feature values
  - Baseline 2: “flatten” adjacency matrix



Case study: brain graph classification

Method Category	Unweighted		Weighted	
	Ordinary	Surprising	Ordinary	Surprising
EAGLE-FLEX	0.6893	0.5499	0.7096	**0.7296**
EAGLE-FIX: 6	0.5114	0.5445	0.6961	**0.7371**
EAGLE-FIX: 8	0.6795	0.5904	**0.7216**	0.7079
EAGLE-FIX: 10	0.5003	0.4989	**0.7032**	0.6807
Full	-	-	0.6681	0.7147
Baselines	Baseline 1:	0.7028	Baseline 2:	0.1099

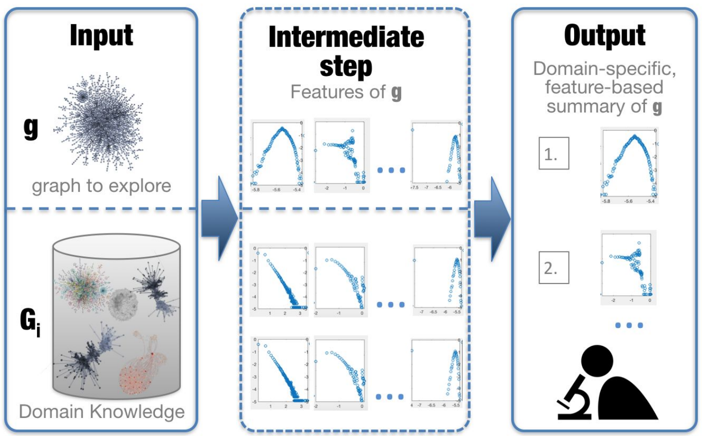
Classification on COBRE: AUC scores per method

Although not designed explicitly for this, features selected by EAGLE can be applied to specific tasks such as classification with at least as good performance.

EAGLE-Flex improves performance by effectively eliminating noise from the data.

Conclusion & Contributions

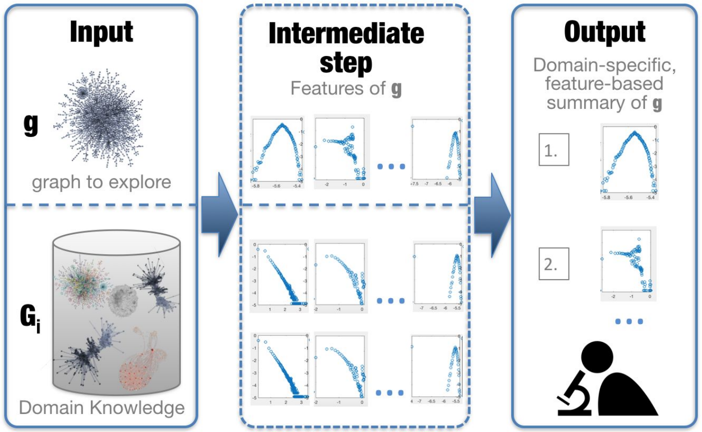
- EAGLE: a **novel graph summarization** technique that *learns an unknown graph from known ones*.
- Informative graph features that satisfy:
  - Diversity
  - Conciseness
  - Domain-specificity
  - Interpretability
  - Efficiency



- Formulation of graph exploration as **constrained optimization**
  - Two efficient solutions: Eagle-Fix and Eagle-Flex.
  - Applications.

Thank you! Questions?

- EAGLE: a **novel graph summarization** technique that *learns an unknown graph from known ones*.
- Informative graph features that satisfy:
  - Diversity
  - Conciseness
  - Domain-specificity
  - Interpretability
  - Efficiency



- Formulation of graph exploration as **constrained optimization**
  - Two efficient solutions: Eagle-Fix and Eagle-Flex.
  - Applications.