On Transductive Classification in Heterogeneous Information Networks

Xiang Li, Ben Kao, Yudian Zheng, Zhipeng Huang

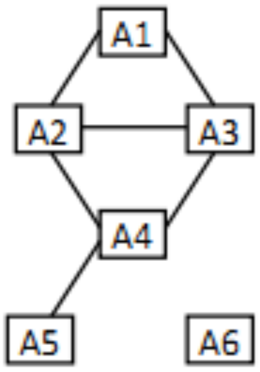
The University of Hong Kong

Outline

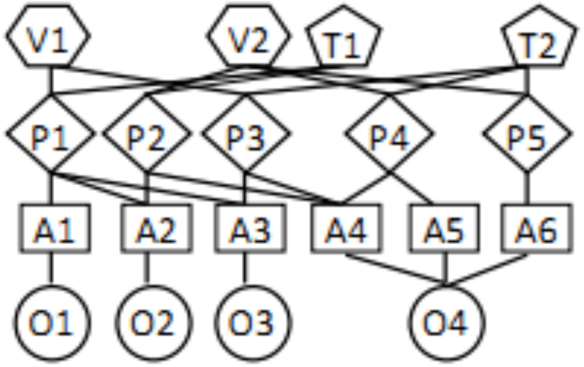
- Introduction
- Experiments and Analysis
- Applications

Introduction

- Homogeneous Information Networks
  - ➢ Objects: entities of the same type
  - ➢ Links: one type of relationships
- Heterogeneous Information Networks (HINs)
  - ➢ Objects: entities of different types
  - ➢ Links: different kinds of relationships



(a) A network of authers



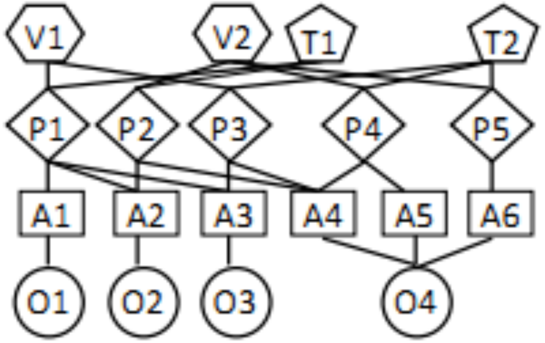
(b) A bibliographic HIN

yago
select knowledge

Freebase

Meta-path

- A ***meta-path*** is a sequence of object types that expresses a relationship between two objects in an HIN
- ***Meta-path*** captures correlation between objects
- e.g., in DBLP network
  - ***APA***: (A1-P1-A2)
  - ***AOA***: (A4-O4-A5)
  - ***APVPA***: (A1-P1-V1-P3-A3)



(b) A bibliographic HIN

Why classification?

- Descriptive labels
  - research area for author
  - genre for movie
- Labeling objects
  - Costly manual effort
  - Incomplete labels (e.g. 75% adventure genre movies in Yago miss the label)

Category of classification

✗ Inductive classification

  - Train a model based on labeled objects

- Transductive classification

  - Utilize “relatedness” between objects to “propagate” labels

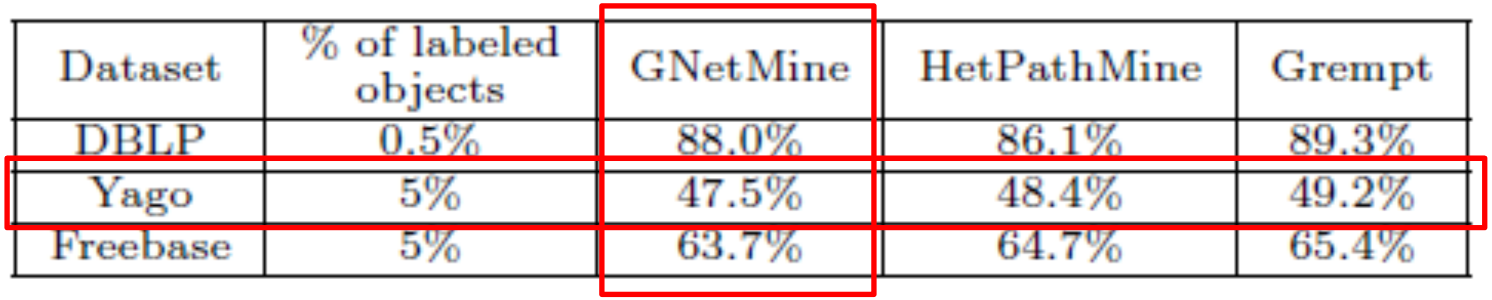
- Relatedness ⇨

Edge relation
Path relation (meta path in HIN)

HINs with scarce labeled data

Two observations

- Cross-sectional study
  - Comparable results on the same task
- Longitudinal study
  - Greatly varied performance over different tasks



Dataset	% of labeled objects	GNetMine	HetPathMine	Grempt
DBLP	0.5%	88.0%	86.1%	89.3%
Yago	5%	47.5%	48.4%	49.2%
Freebase	5%	63.7%	64.7%	65.4%

Table 1: Accuracies of transductive classifiers

Summary

- For transductive classification in HINs:
  - □ Marginal benefits in fine tuning the algorithms
  - □ Latent factors influence its success

Classification tasks

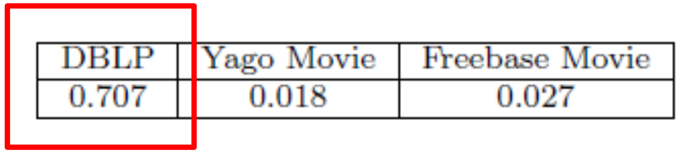
Dataset	Task	Description	Links	Label set	Meta path set
DBLP	Classify authors	14,376 papers (P) 20 venues (V) 14,475 authors (A) 8,920 terms (T)	P-A P-V P-T	DB DM AI IR	APA, APAPA, APVPA, APTPA
Yago Movie	Classify movies	1,465 movies (M) 4,019 actors (A) 1,093 directors (D) 1,458 writers (W)	M-A M-D M-W	horror action adventure	MAM, MDM, MWM, MAMAM, MDMDM, MWMWM
Freebase Movie	Classify movies	3,492 movies (M) 33,401 actors (A) 2,502 directors (D) 4,459 producers (P)	M-A M-D M-P	faction adventure crime	MAM, MDM, MPM, MAMAM, MDMDM, MPMPM

Connectivity assumption

- For any two objects, if they are highly connected (by links or paths), they are more likely to share the same label

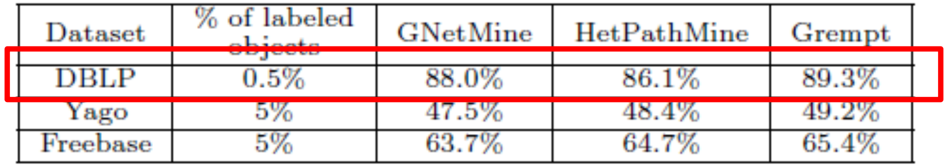
Question 1: Does the connectivity assumption generally hold?

- **NetClus**: cluster objects based on network structure
- Compare **NetClus-induced clusters** with **true-label-induced clusters**
- The higher the similarity, the more likely highly connected objects share the same label, the better performance of transductive classifers



DBLP	Yago Movie	Freebase Movie
0.707	0.018	0.027

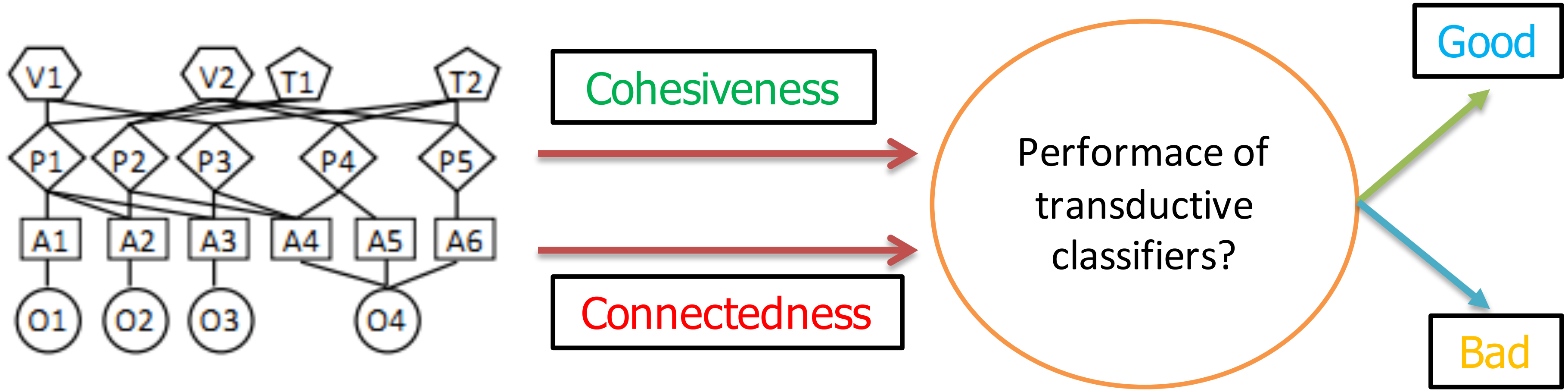
Table 3: Similarity (NMI) of $\mathcal{C}_{\hat{L}}$ and $\mathcal{C}_{NetClus}$



Dataset	% of labeled objects	GNetMine	HetPathMine	Grempt
DBLP	0.5%	88.0%	86.1%	89.3%
Yago	5%	47.5%	48.4%	49.2%
Freebase	5%	63.7%	64.7%	65.4%

Table 1: Accuracies of transductive classifiers

Question 2: When will transductive classifiers work in an HIN?



Cohesiveness

Connectedness

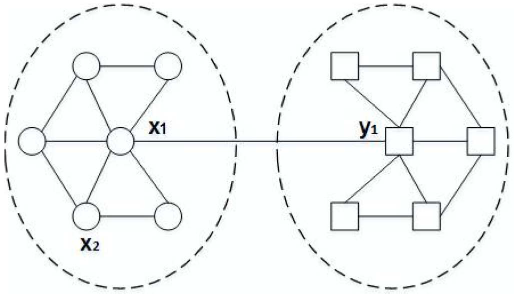
Performace of transductive classifiers?

Good

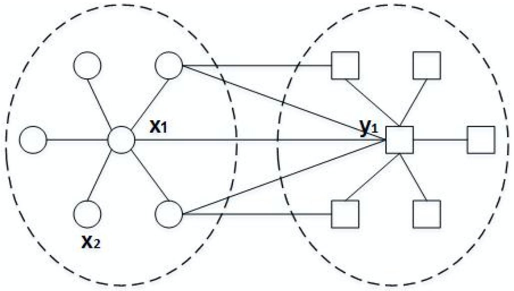
Bad

Cohesiveness

intra-cluster edges are more
inter-cluster edges are fewer



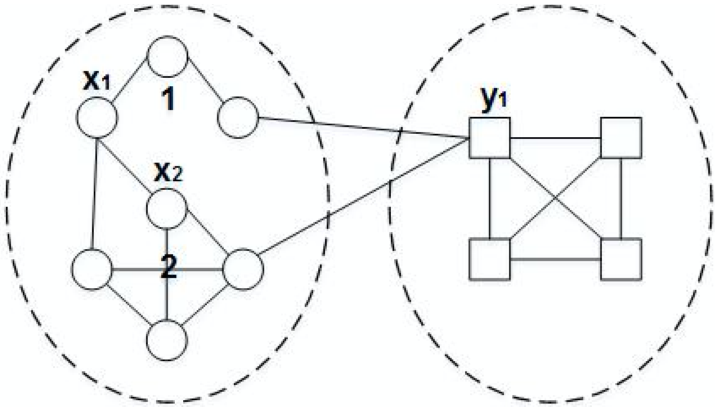
(a). A cohesive network



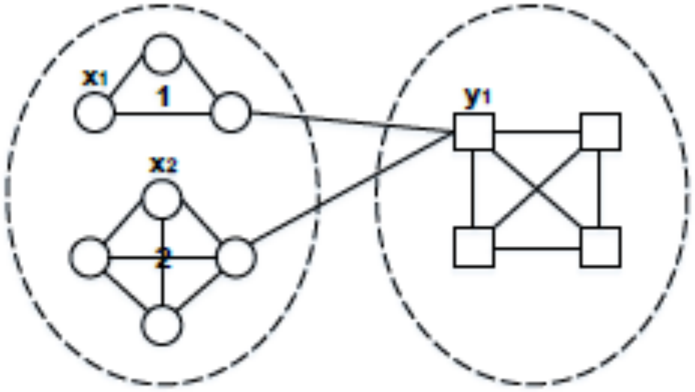
(b). A non-cohesive network

Connectedness

- Intuitively, an HIN is highly ***connected*** if objects of the same label exhibit strong connectivity



(a). A connected network



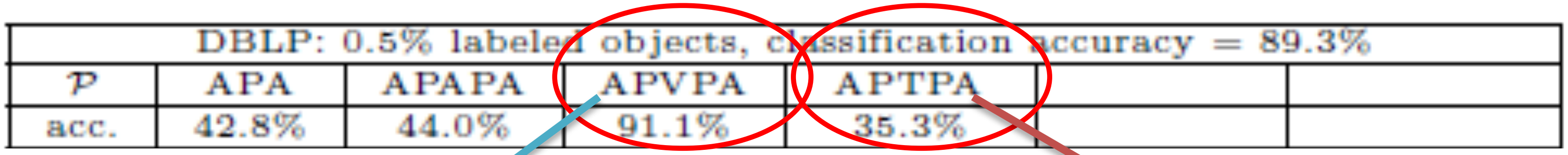
(b). A less connected network

How are cohesiveness and connectedness correlated with classification accuracy?

- DBLP has much larger cohesiveness ϒ and connectedness ψ -> higher classification accuracy

$\Upsilon_{APVPA} = 0.393$
$\psi_{APVPA} = 1.0$

$\Upsilon_{APVPA} = 0.016$
$\psi_{APVPA} = 1.0$



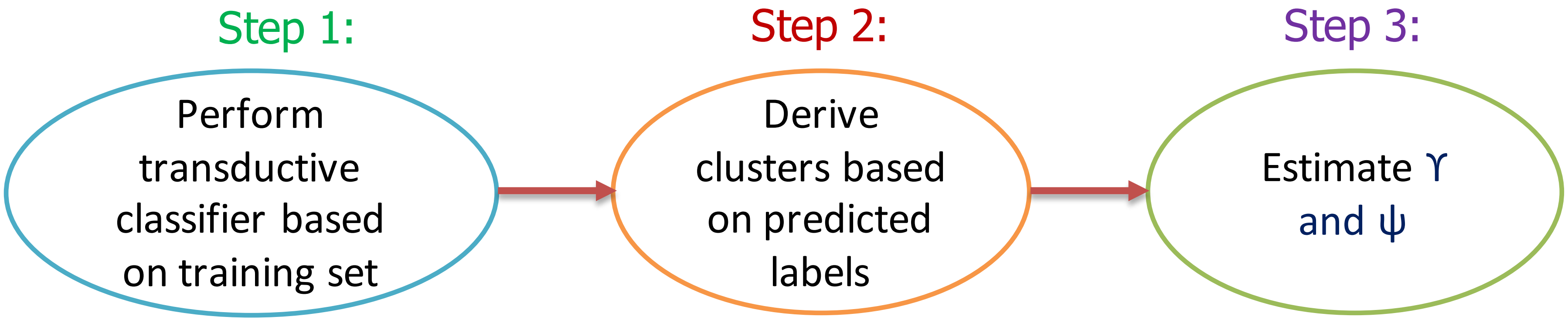
DBLP: 0.5% labeled objects, classification accuracy = 89.3%						
$\mathcal{P}$	APA	APAPA	APVPA	APTPA		
acc.	42.8%	44.0%	91.1%	35.3%		

Authors publish papers in the same conference

Authors publish papers using same keyword

Estimate cohesiveness ϒ and connectedness ψ

- In fact, we only have a small set of labeled objects
- To estimate cohesiveness ϒ and connectedness ψ:



Step 1:

Perform transductive classifier based on training set

Step 2:

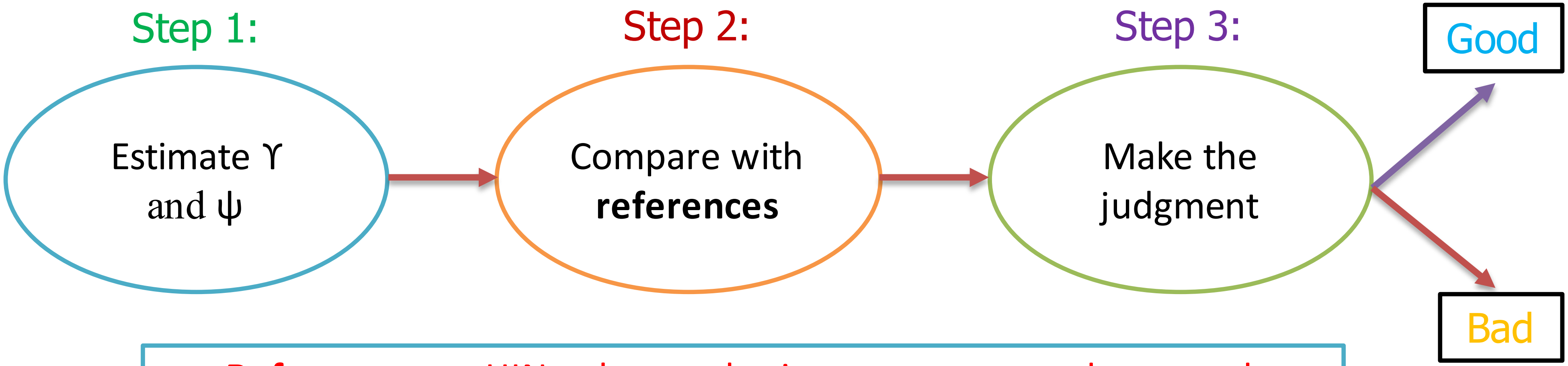
Derive clusters based on predicted labels

Step 3:

Estimate ϒ and ψ

Black-box tester

- Recommend whether transductive classification should be applied
- The procedures:



References are HINs whose cohesiveness, connectedness, and performance of transductive classifiers have already been known

Active learner (ALCC)

- **Quality score**: QS = estimated ϒ * estimated ψ
- Each iteration selects Ns objects leading to the largest improvement in QS
- Iteration repeats until budget B exhausts

Observations

- DBLP: estimated ϒ and estimated ψ close to true ones
- Yago Movie and Freebase Movie :
  - Estimated ψ is close to true ψ
  - Estimated ϒ is overestimated

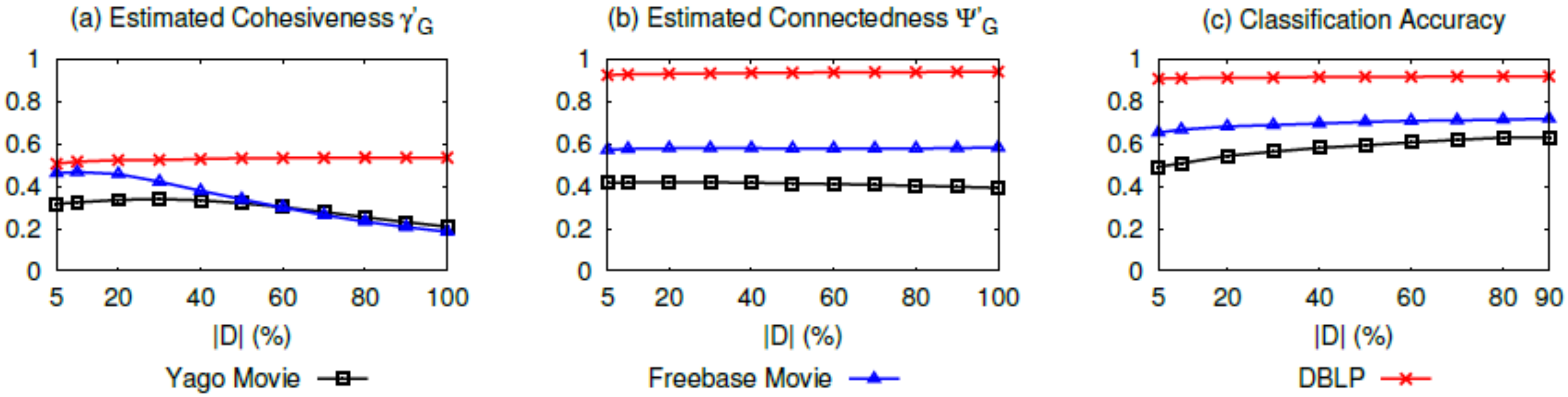


Figure 5: Estimating cohesiveness, connectedness, and classification accuracy of 3 HIN classification tasks

Case studies

Dataset	Task	Description	Links	Label set	Meta path set
TV	Classify series	2,913 series (S) 652 directors (D) 685 writer (W) 151 TV programs (P)	S-D S-W S-P	comedy-drama soap opera police procedural.	SDS, SWS, SPS, SDSDS, SWSWS, SPSPS
Game	Classify games	4,095 games (G) 1,578 publishers (P) 2,043 developers (D) 197 designers (S).	G-P G-D G-S	action adventure strategy	GPG, GDG, GSG, GPGPG, GDGDG, GSGSG

Results of Black-box tester

- Training set: 15% objects

Dataset	(estimated ϒ, estimated ψ)	(true ϒ, true ψ)	Classification Accuracy
TV	(0.749, 0.836)	**(0.887, 0.889)**	**94.3%**
Game	(0.342, 0.254)	(0.250, 0.297)	34.2%

Dataset	true ϒ	true ψ	Transductive classifier performance
DBLP	0.536	0.942	good
Yago Movie	0.209	0.393	bad
Freebase Movie	0.185	0.584	bad

Table: References

Active learning

1. Random performs the worst
2. ALGE [global entropy] is generally better than US [Local entropy]
3. ALCC always performs the best

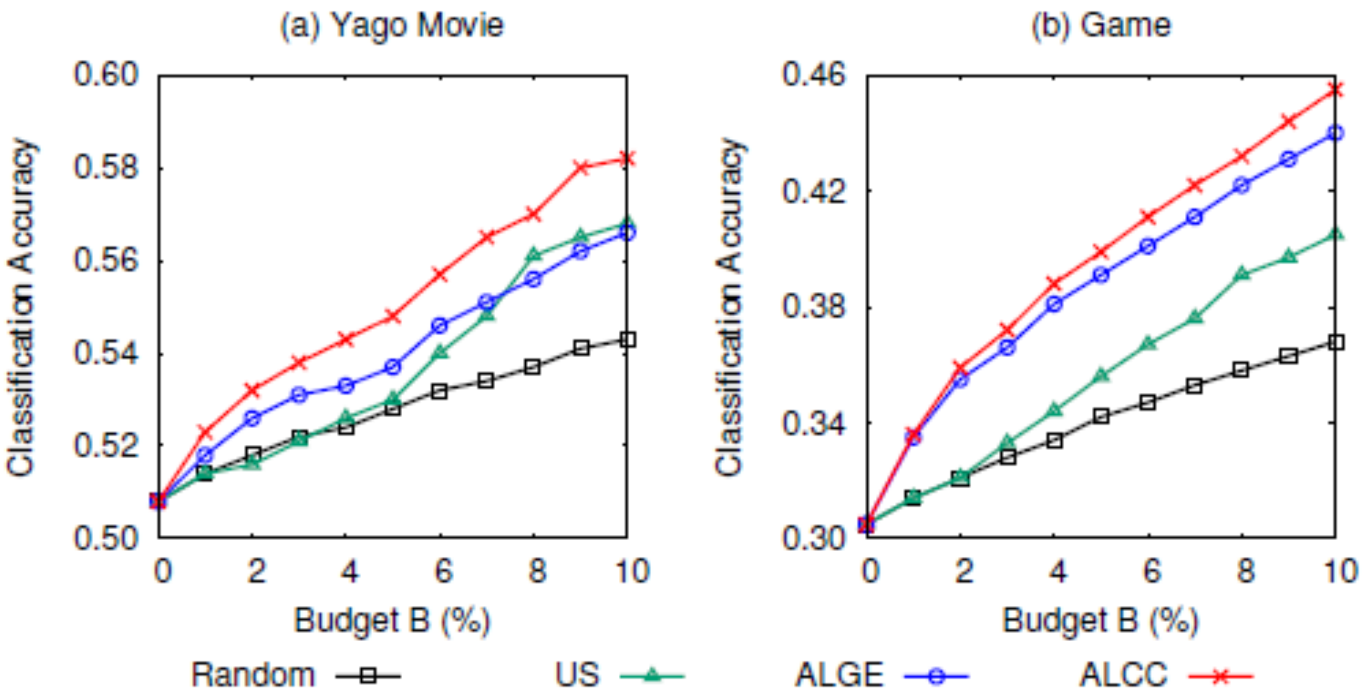


Figure 7: Active learner comparison

Conclusion

- Provide a thorough analysis to tranductive classification in HINs
- Identify two influential factors
- Design a useful black-box tester
- Propose an effective active learning method

Thank you!