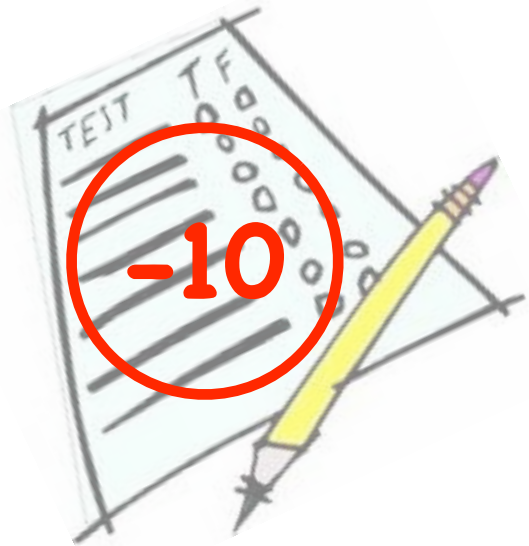
Evaluation Strategies for Network Classification

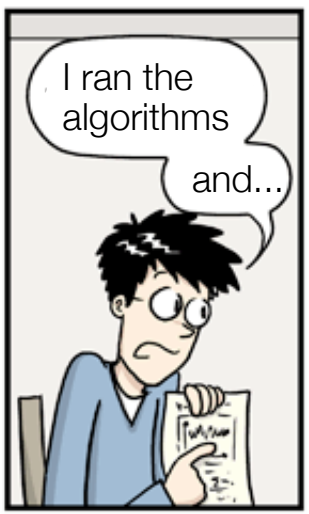
Jennifer Neville
Departments of Computer Science and Statistics
Purdue University

(joint work with Tao Wang, Brian Gallagher, and Tina Eliassi-Rad)

Given two learning algorithms A and B and a dataset of size S from a domain D...

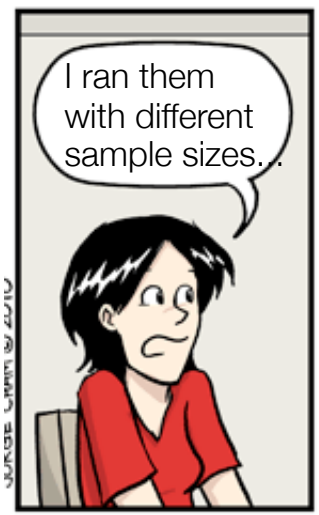
which ***algorithm*** *will produce more accurate classifiers when* ***trained*** *on other* ***datasets*** *of size S drawn from D?*

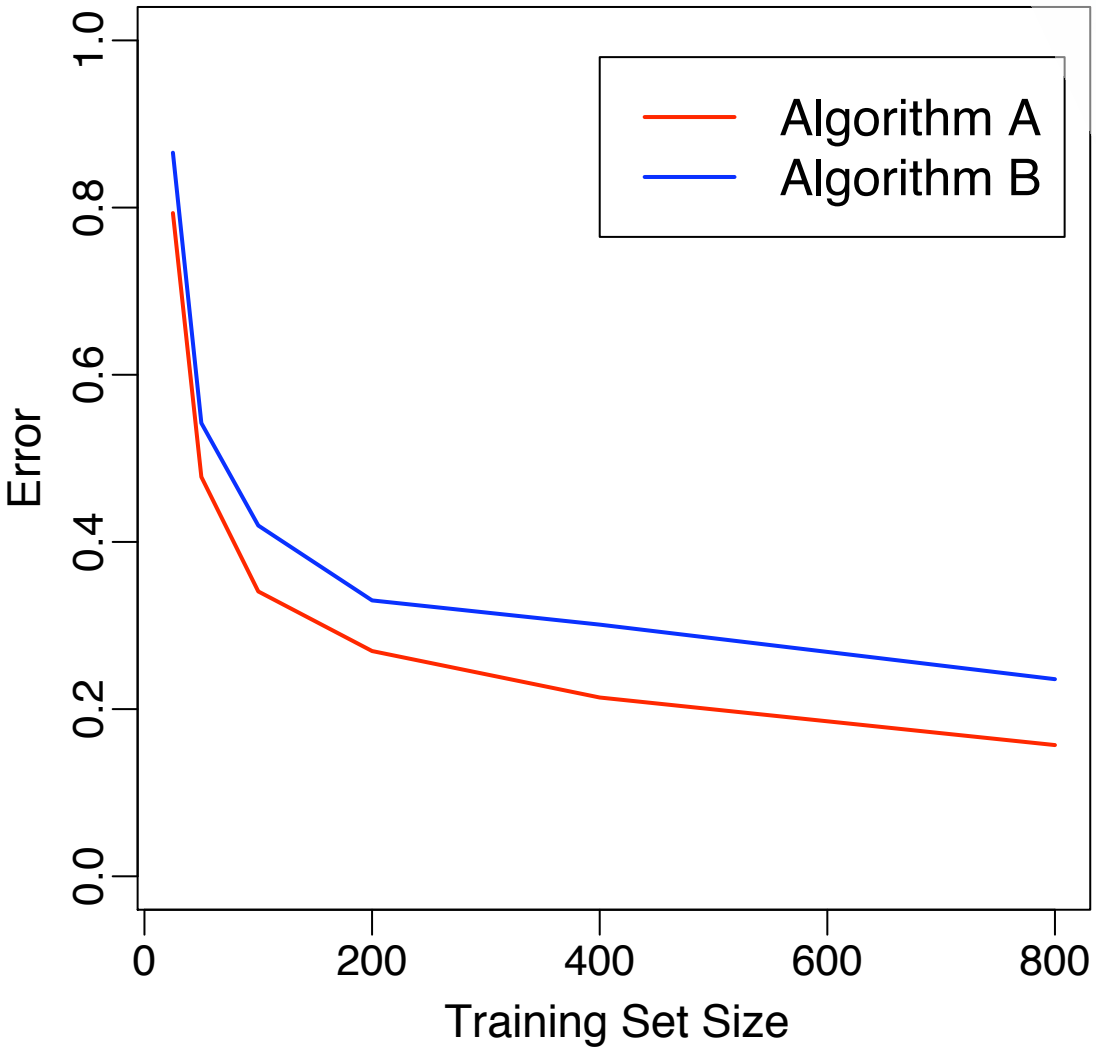


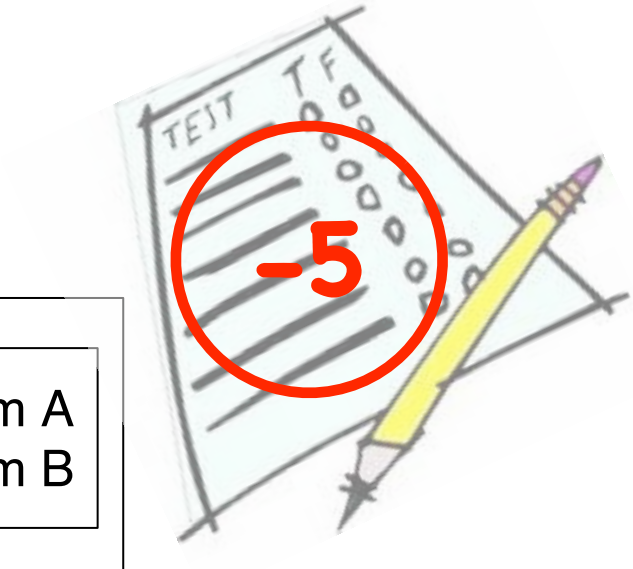


Algorithm A: Error rate=0.19
Algorithm B: Error rate=0.21

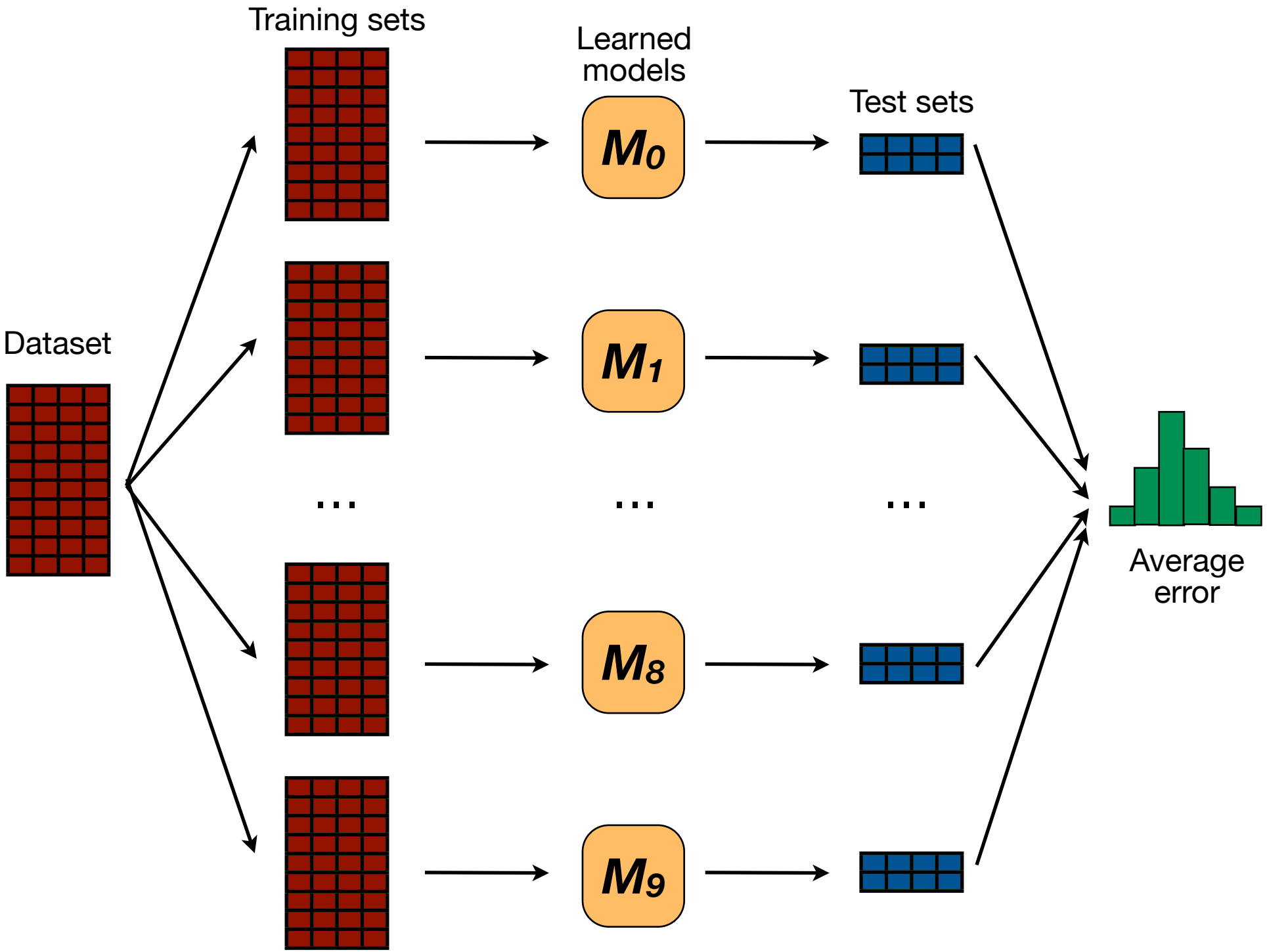
Algorithm A is better than Algorithm B!

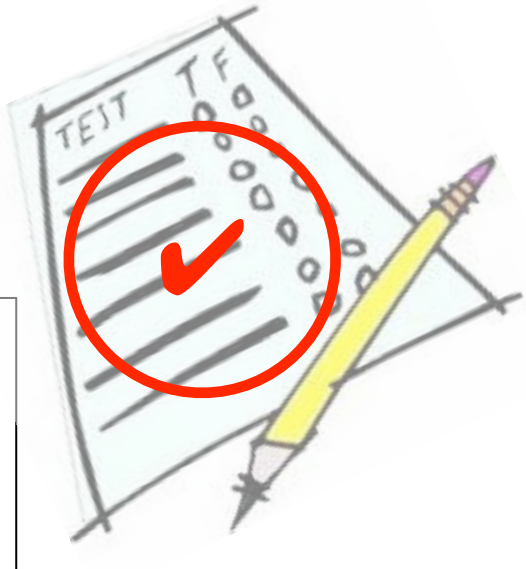


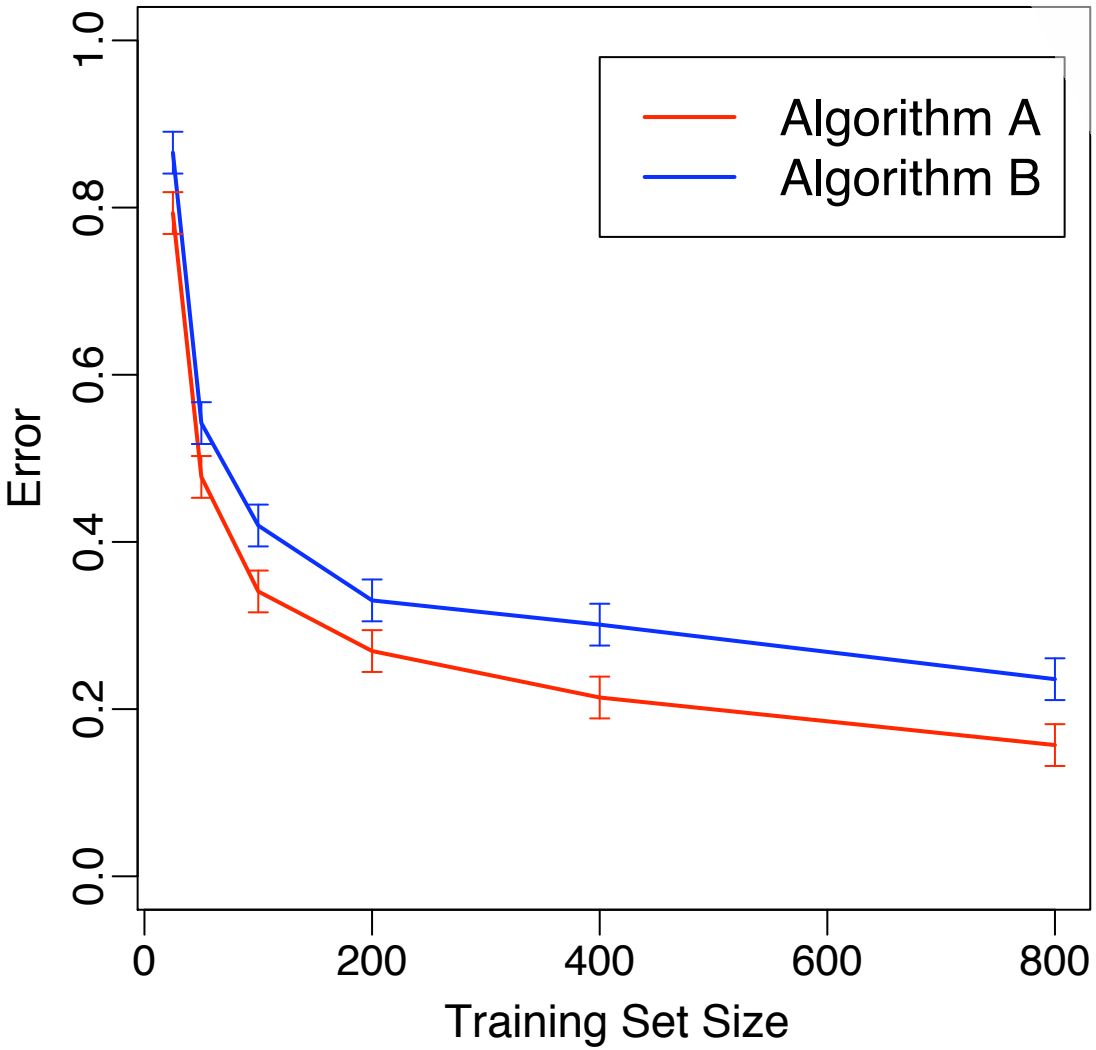




Algorithm A is better than Algorithm B!


Training sets
Learned models
Test sets
Dataset
M_0
M_1
M_8
M_9
Average error



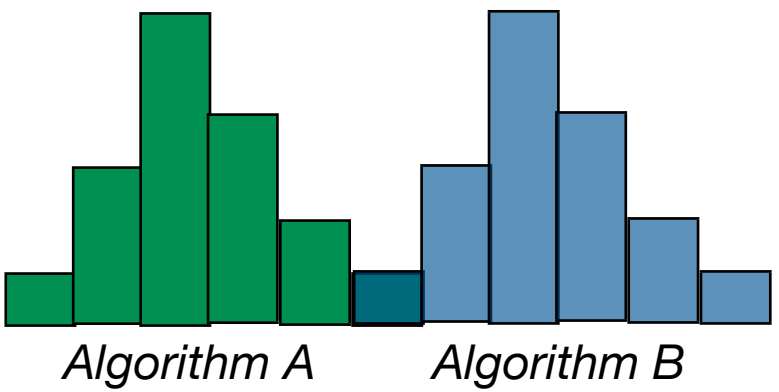


Algorithm A is better than Algorithm B!

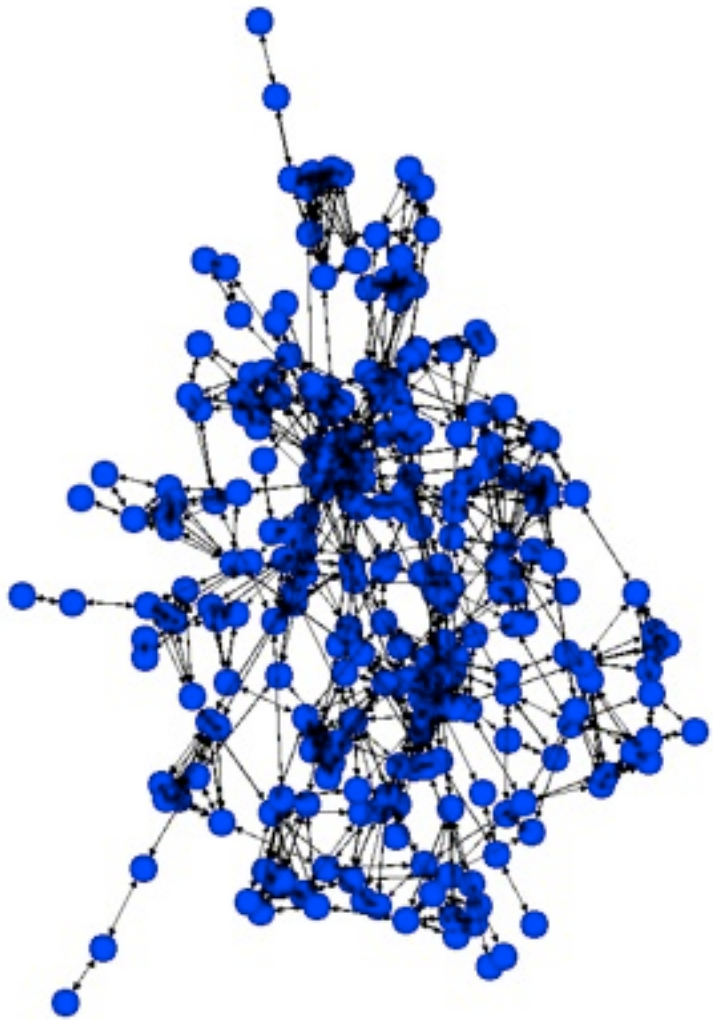
To assess algorithm performance,
you conduct a **hypothesis** test
(either implicitly or explicitly)

Comparison of algorithm performance

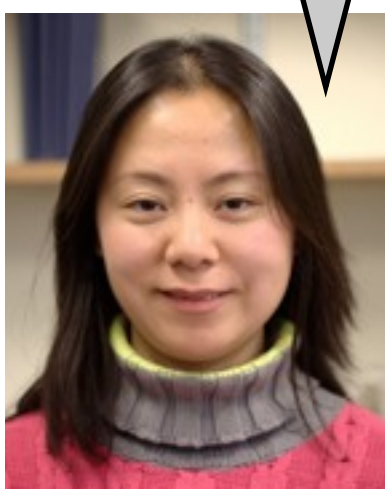
- **Sampling**
  - How to partition or sample available data into training and test sets?
  - **k-fold cross-validation** is often used
- **Significance test**
  - Is the observed difference in performance **significantly greater** than what would be expected by random chance?
  - Null hypothesis (H_0): Algorithm performance rates are drawn from the same distribution
  - Two-sample **t-test** is often used

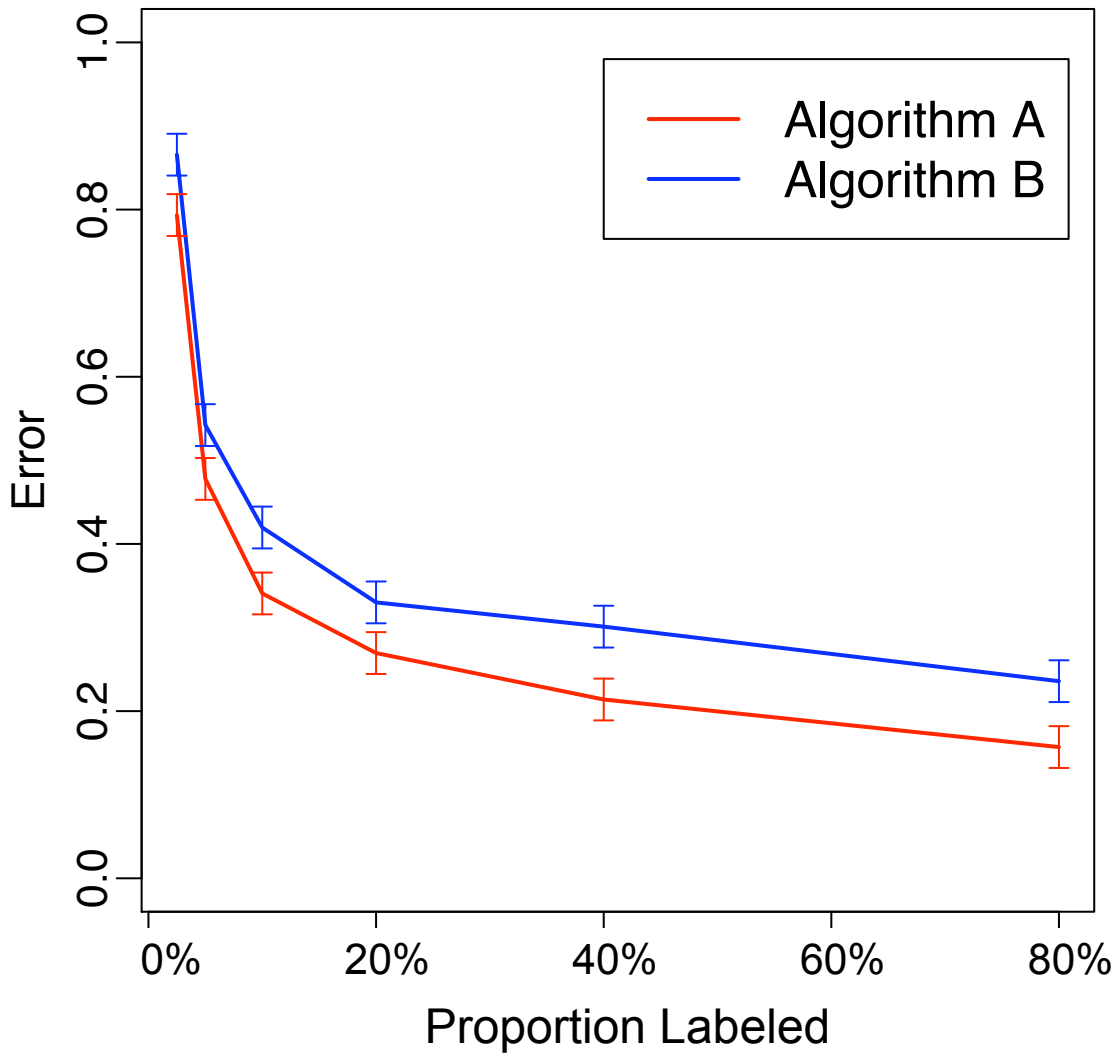


Now what if I told you the dataset was a network?



I ran collective inference at different labeling proportions...

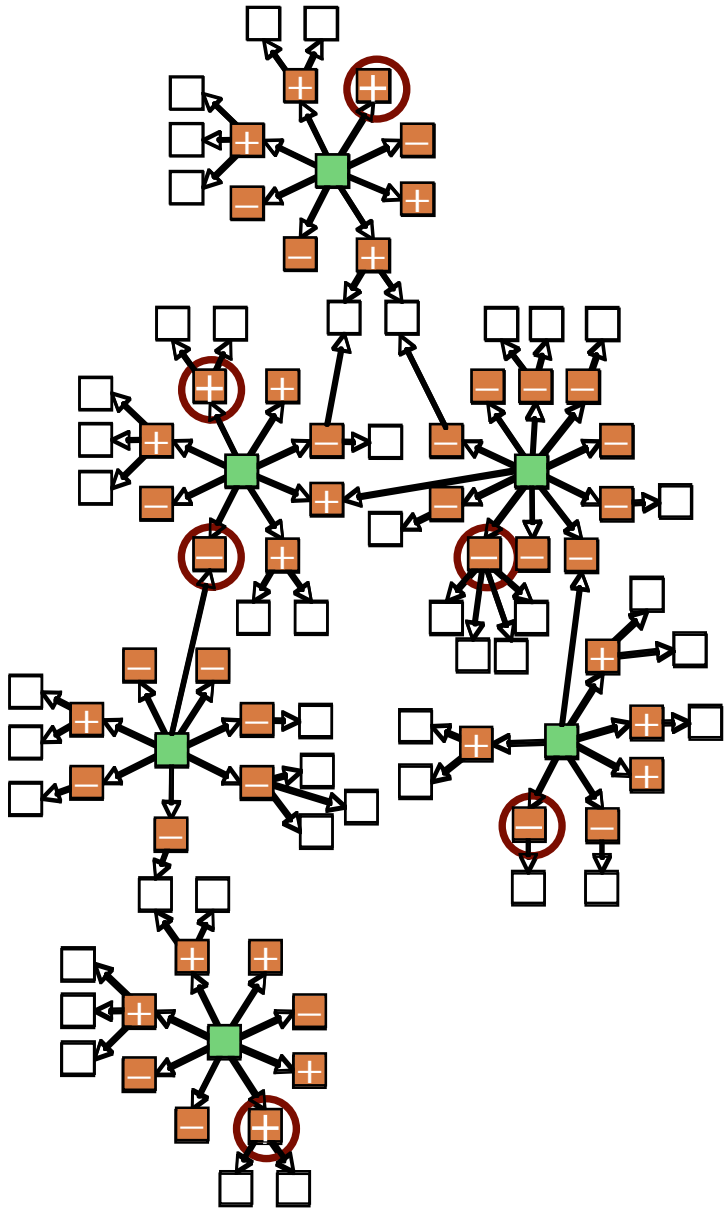




Is Algorithm A better than Algorithm B?

Relational learning and collective inference

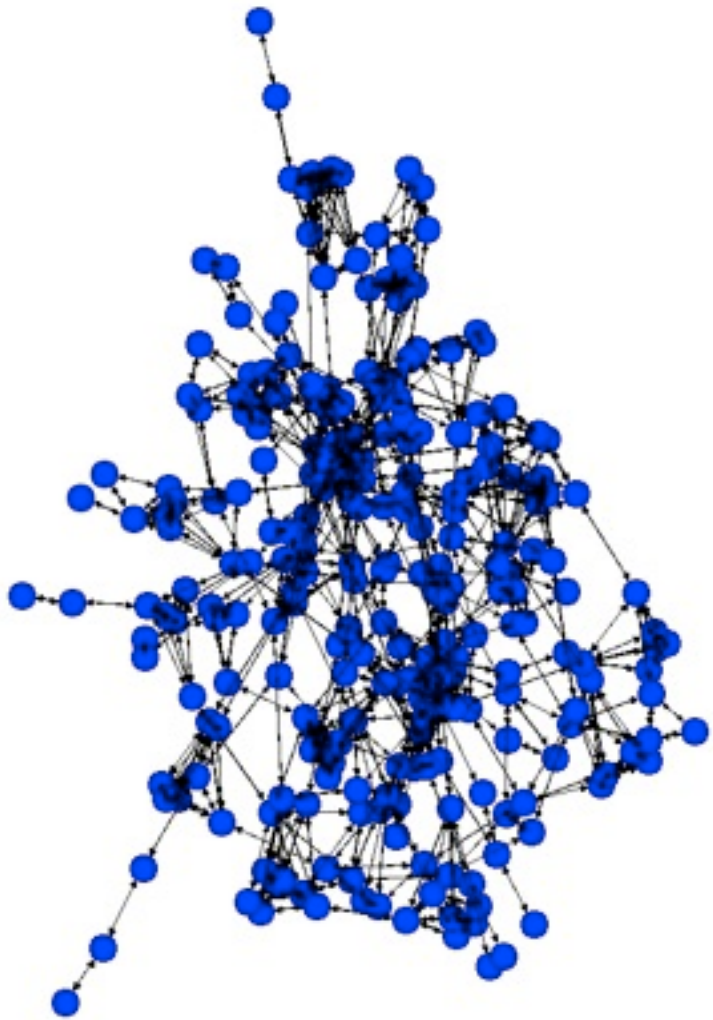
- **Within-network learning**
  - Estimate model from a partially-labeled network
  - Apply learned model to predict the class labels in the remainder of the network (i.e., the unlabeled nodes)
- **Across-network learning**
  - Estimate models from a fully-labeled network
  - Apply learned model to a partially-labeled network, predict class labels for unlabeled nodes



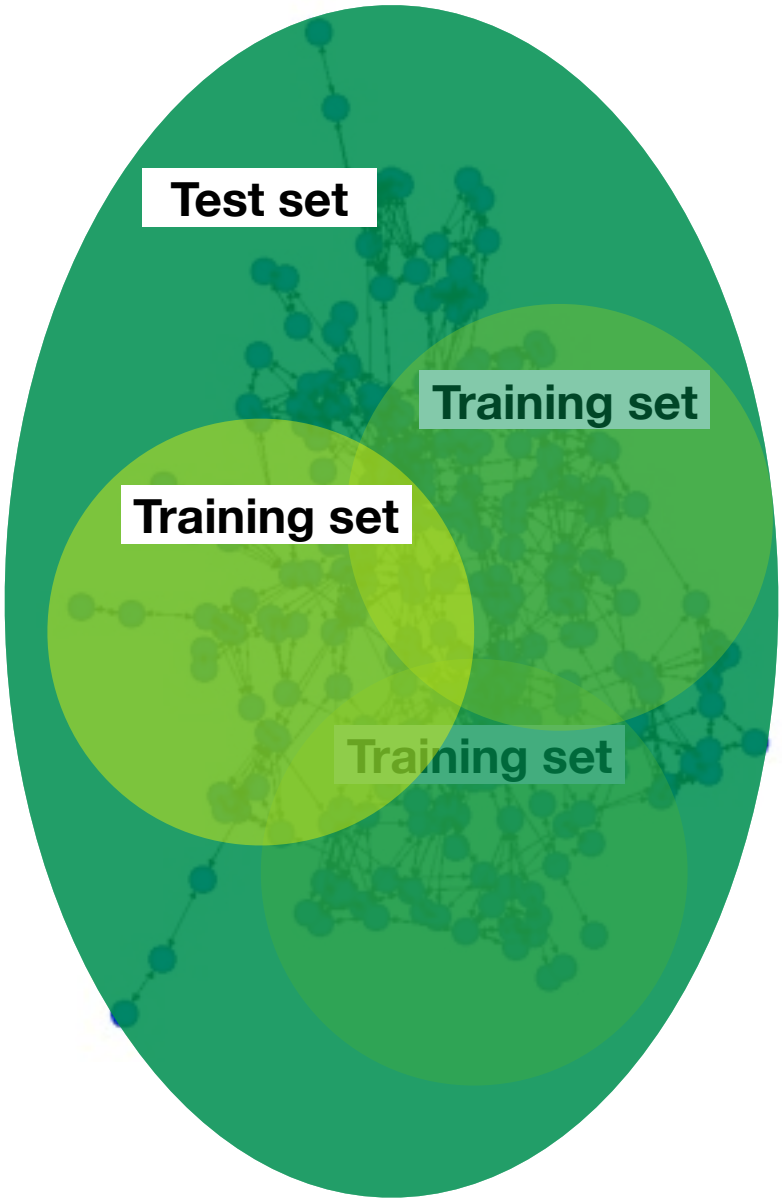
Survey of methodology in 23 recent research papers on relational learning

Resampling Procedure		**Systematic Variation of % Labeled**	
Cross validation	14	No	13
Simple random	8	Yes	10
Controlled random	3	Within-network	8
Snowball sampling	2	Across network	2
Temporal resampling	2		
Statistical Test		**Number of Resampling Folds**	
t-test	10	10	14
StDev/Var/StErr	6	<10	7
None	6	>10	2
Wilcoxon signed rank	2	Unspecified	2
Within vs. Across Network Classification		**Performance Measure**	
Within-network	13	Accuracy	14
Across network	8	AUC	10
Unspecified	6	Precision/Recall/F1	1

How do we sample
a single network?



How do we sample networks?



Common approach

Use **repeated random sampling** to create multiple sets of labeled/ unlabeled nodes

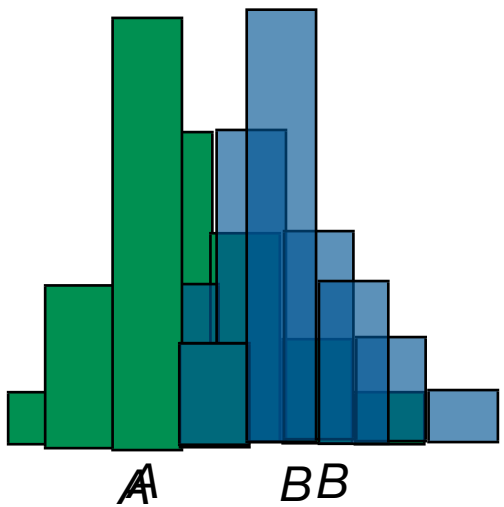
Survey of methodology in 23 recent research papers on relational learning

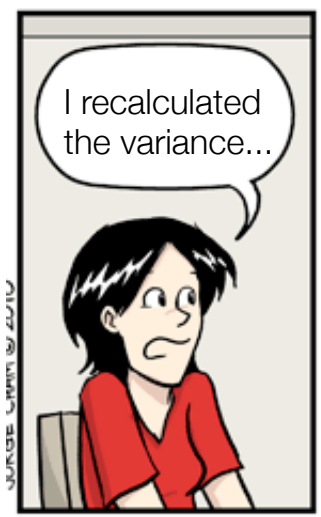
Resampling Procedure		**Systematic Variation of % Labeled**	
Cross validation	14	No	13
Simple random	8	Yes	10
Controlled random	3	Within-network	8
Snowball sampling	2	Across network	2
Temporal resampling	2		
Statistical Test		**Number of Resampling Folds**	
t-test	10	10	14
StDev/Var/StErr	6	<10	7
None	6	>10	2
Wilcoxon signed rank	2	Unspecified	2
Within vs. Across Network Classification		**Performance Measure**	
Within-network	13	Accuracy	14
Across network	8	AUC	10
Unspecified	6	Precision/Recall/F1	1

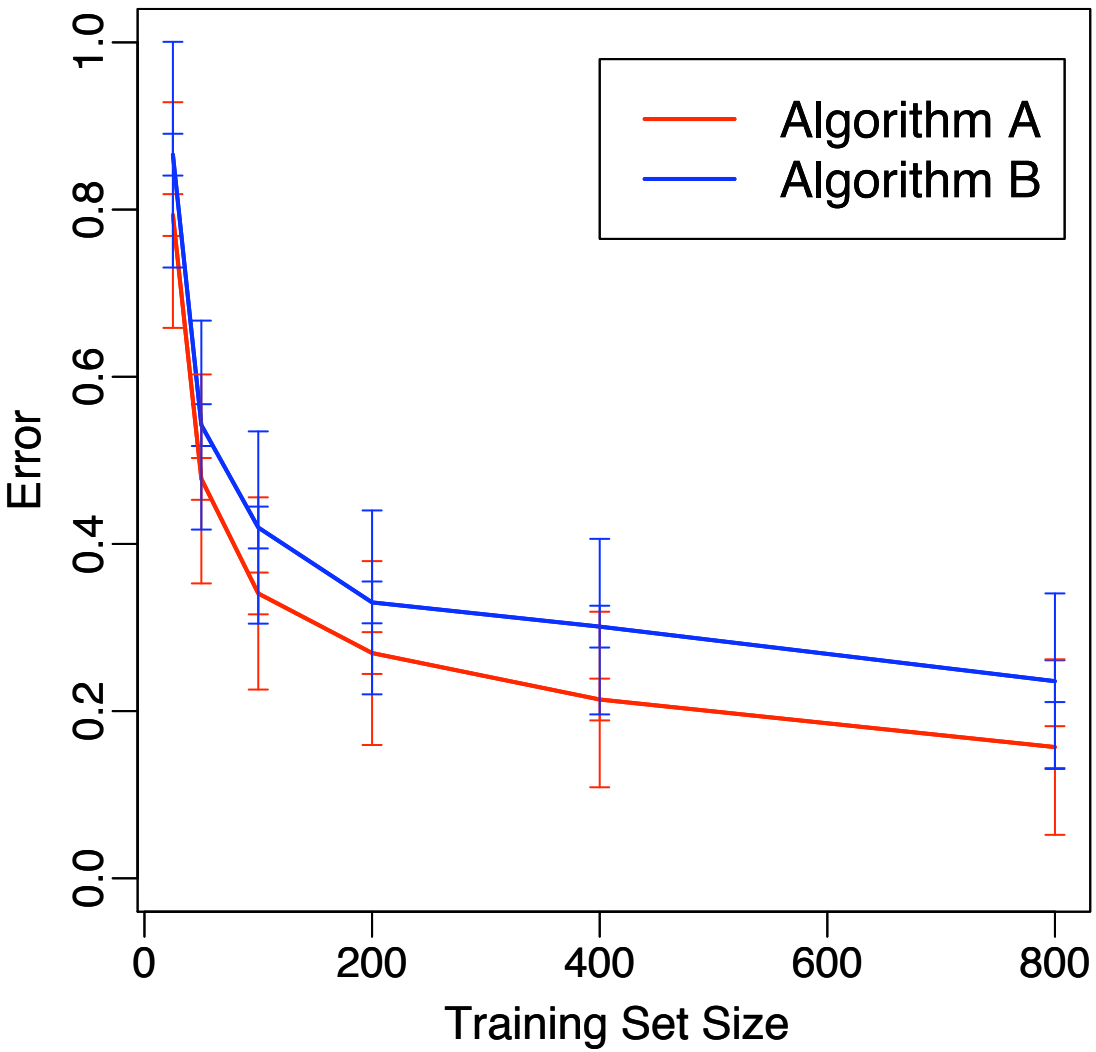
How does simple random sampling affect classifier evaluation?

Overlapping test sets can cause bias

- T-test results are **biased** if performance is estimated from **overlapping** test sets *(Dietterich'98)*
  - Overlapping samples leads to underestimation of variance... which increases the probability of Type I error
  - **Recommendation**: Use cross-validation to eliminate dependencies between test sets





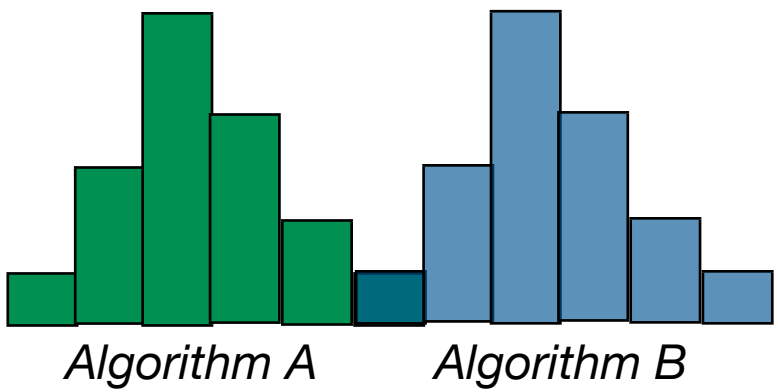


Is Algorithm A better than Algorithm B?

Does the use of resampling affect network domains?

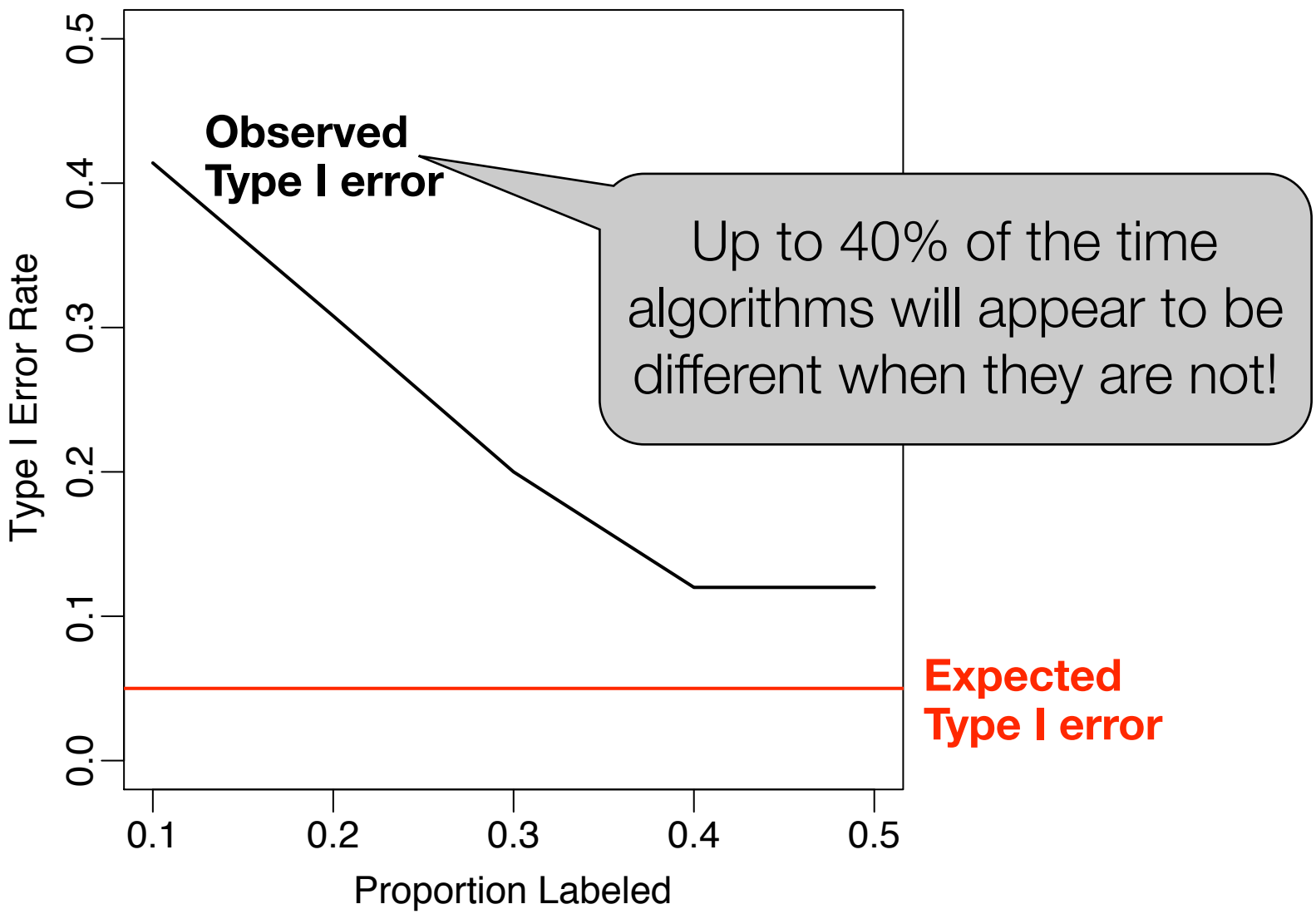
Aspects of hypothesis tests

- **Type I error:**
  - Reject the null hypothesis when it is true (false positive)
  - *Conclude algorithms are different when they are not*



Evaluation of paired t-test on network data *(ICDM'09)*

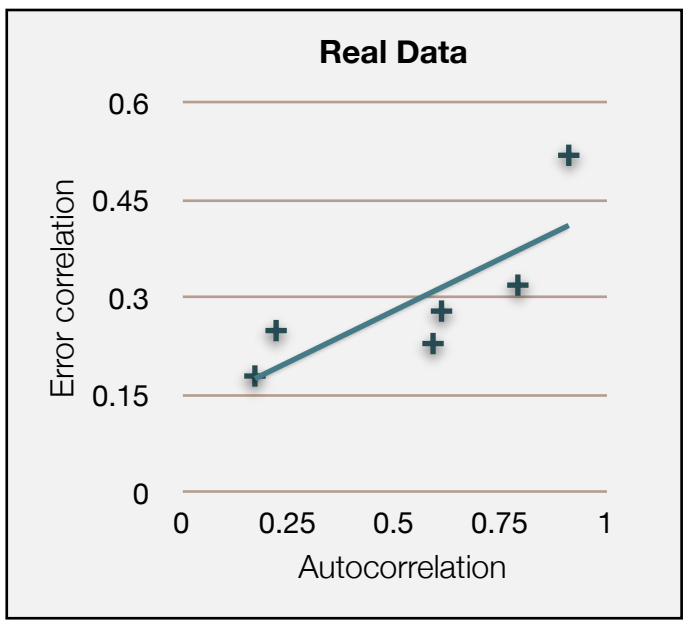
Type I error: *Incorrectly conclude that algorithms are different when they are not*



Network characteristics that lead to bias

- **Training and test set sizes are dependent**
  - As the proportion of labeled data decreases, the size of the test set increases, which... increases the overlap between test sets

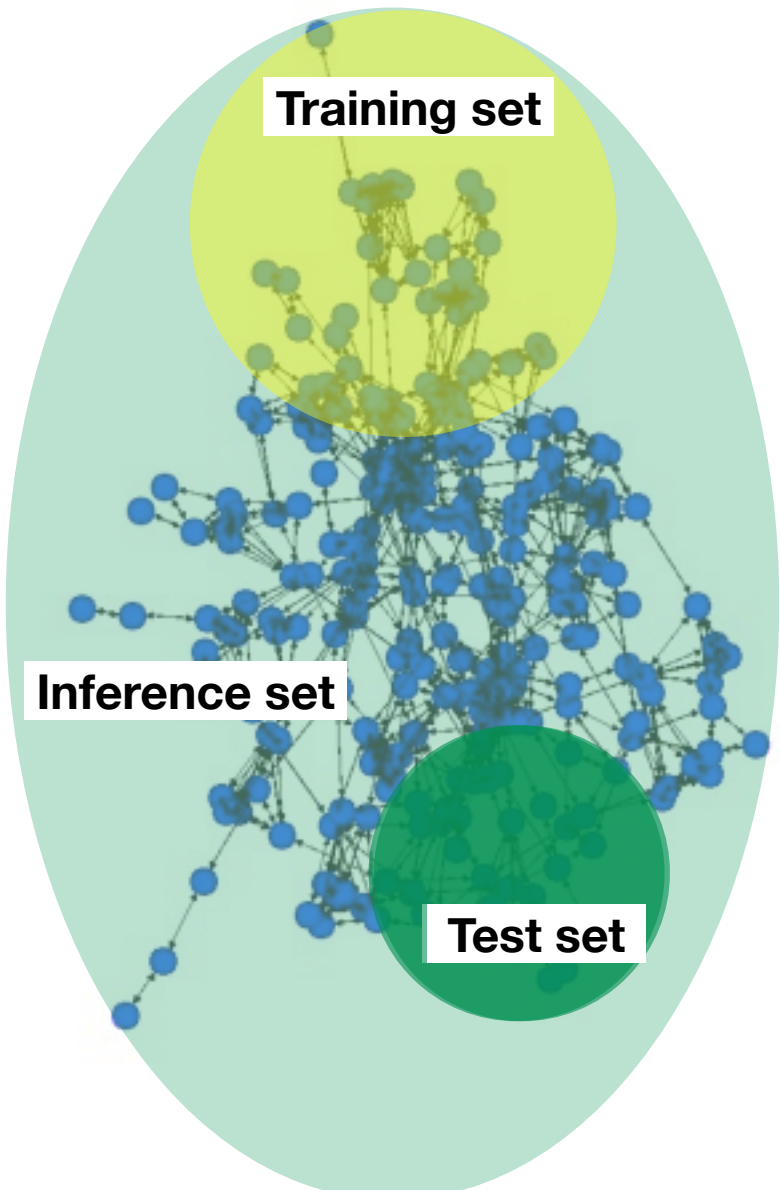
- **Network instances are not independent**
  - Dependencies among instances leads to correlated errors
  - Correlated error increases the variance of algorithm performance



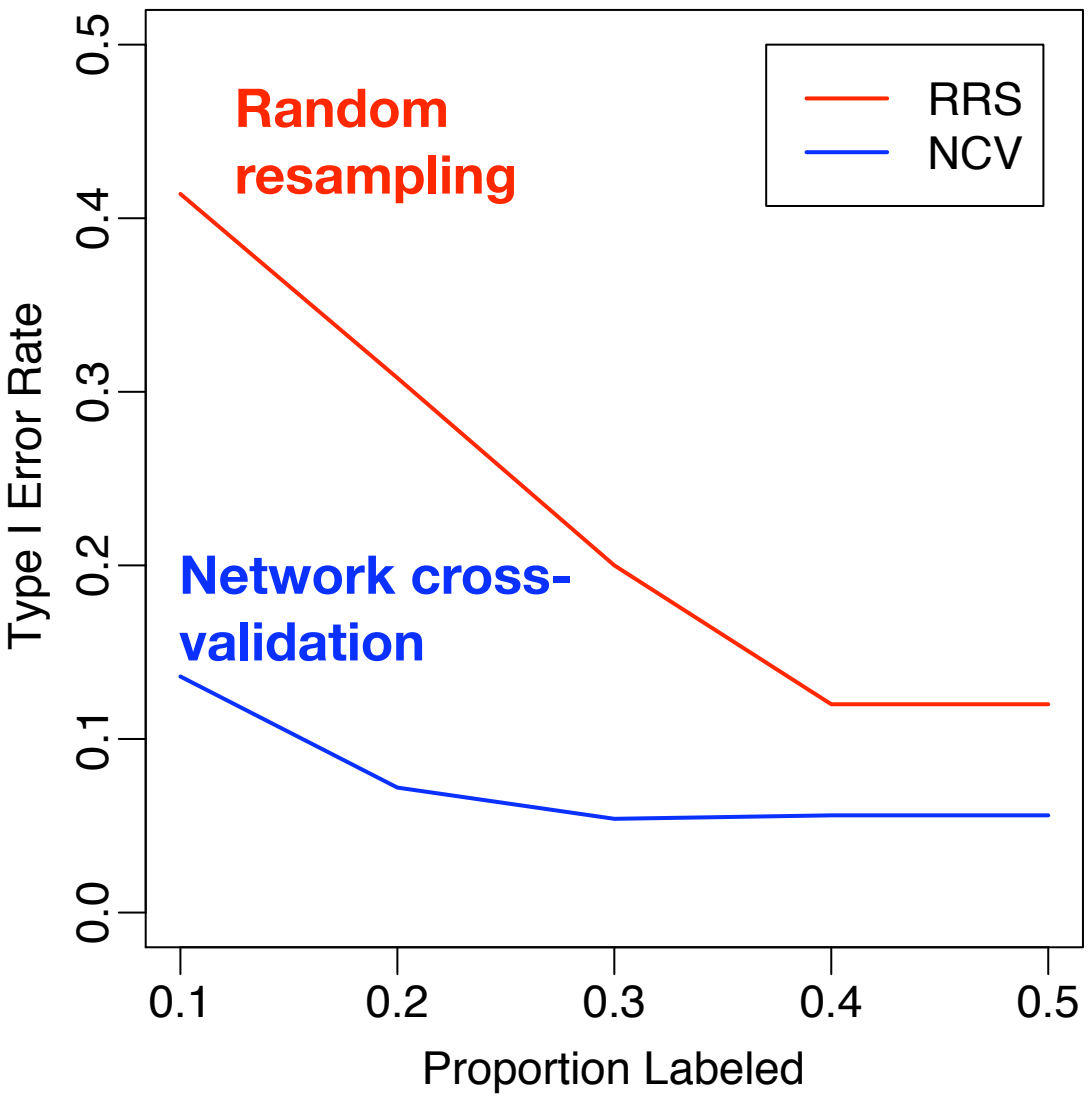
Can we use cross-validation to adjust for bias in network classification?

Network cross-validation *(ICDM'09)*

- Use k-fold cross-validation to select disjoint **test sets** of size N/k
- From remaining N(k-1)/k data randomly select labeled **training set** of appropriate size (e.g., for p% labeled, select p·N instances to label as the training set)
- Add all unlabeled instances to the **inference set** (e.g., network = training set + inference set)
  - Run collective inference over entire inference set to make predictions
  - But only evaluate accuracy of predictions on disjoint test sets



NCV reduces Type I error



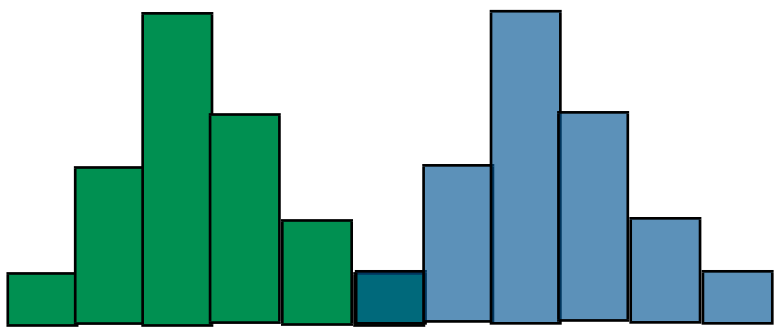
Data: AdHealth dataset, six middle- and high-school social networks

Task: Predict whether a student smokes or not

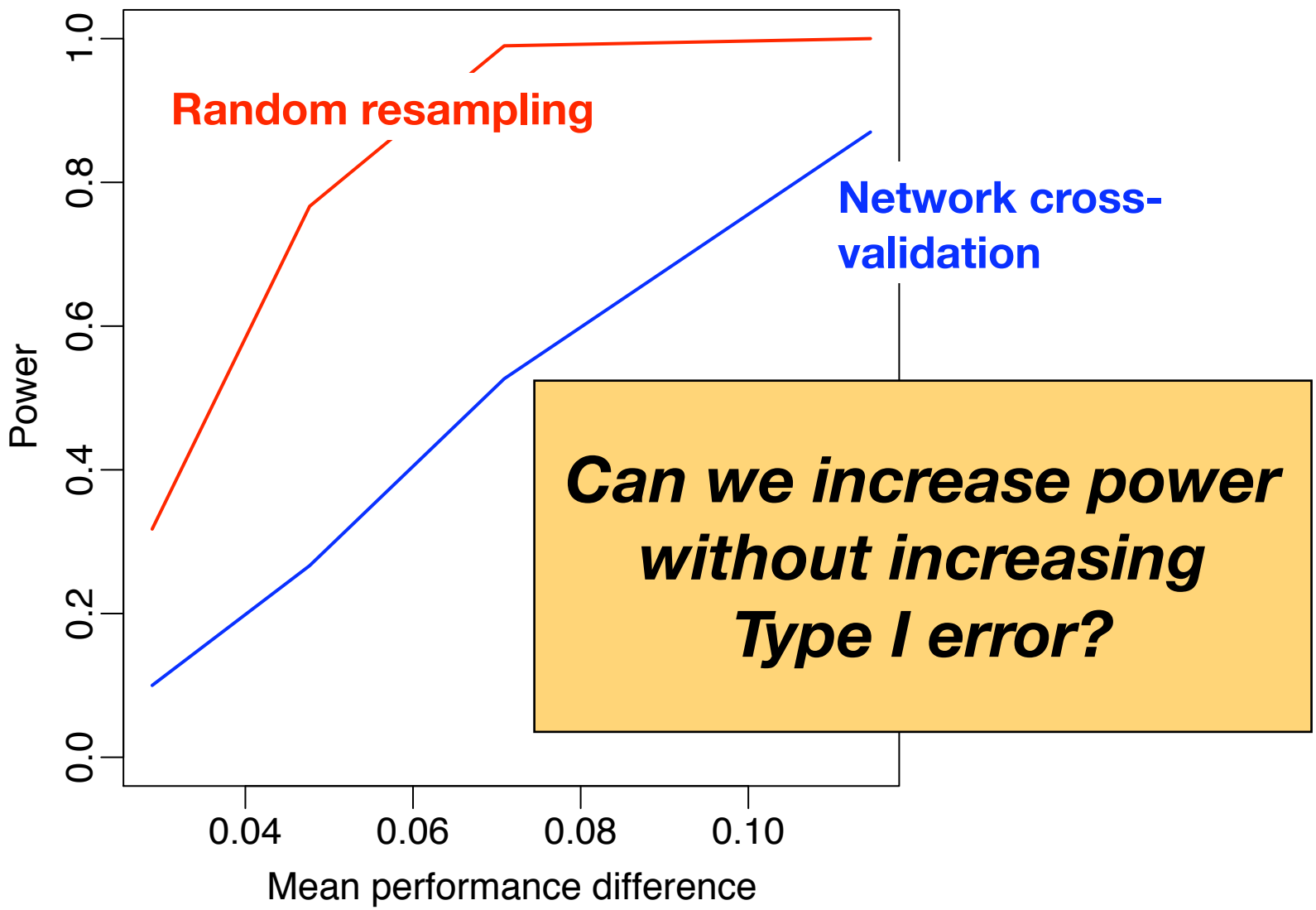
Models: Compare wvRN and nBC algorithms
(Macskassy & Provost JMLR'07)

Aspects of hypothesis tests

- **Type I error:**
  - Reject the null hypothesis when it is true (false positive)
  - *Conclude algorithms are different when they are not*
- **Type II error:**
  - Accept the null when it is false (false negative)
  - *Conclude algorithms are equivalent when they are not*
- **Statistical power:**
  - 1- Type II error (true positive)
  - *Rate at which algorithms are identified as different when they are*



NCV results in decreased statistical power



Can we increase power without increasing Type I error?

NCV has lower statistical power due to the use of smaller (disjoint) test sets.

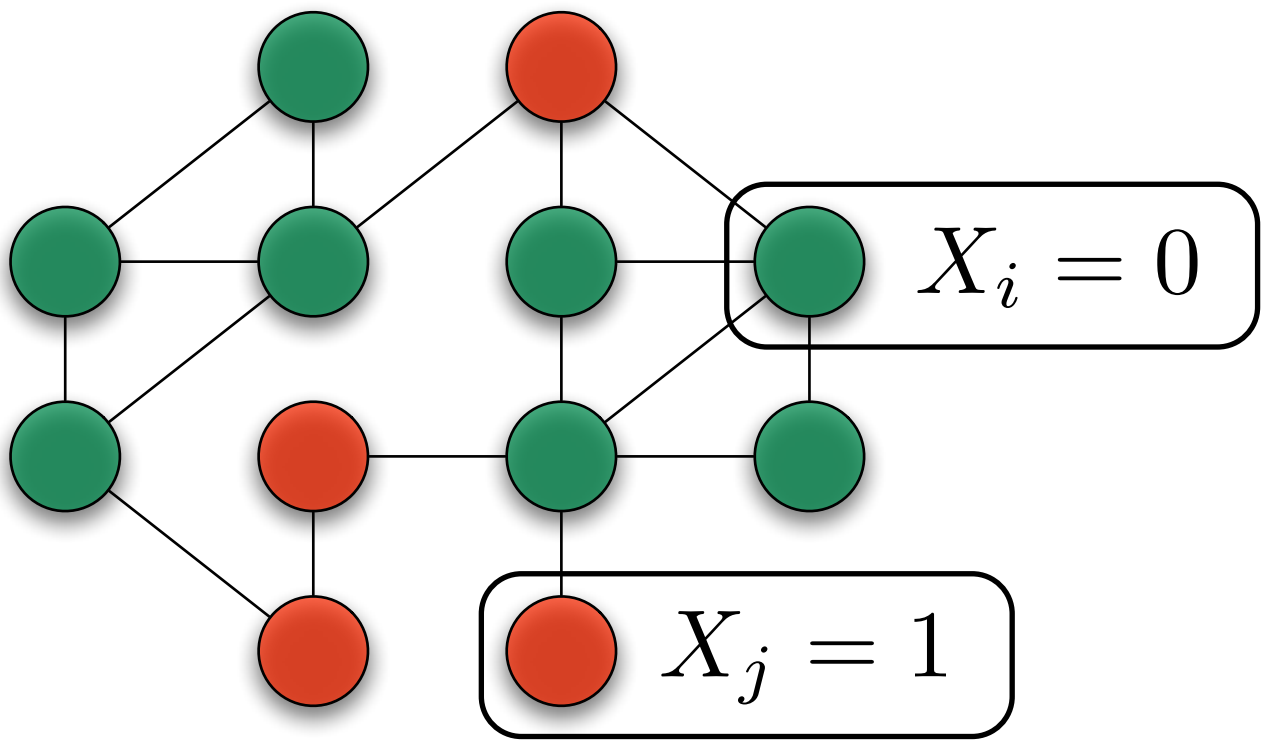
Resampling uses larger (overlapping) test sets that are correlated... but not perfectly correlated.

Can we exploit larger effective sample sizes in resampling by removing effects of overlap?

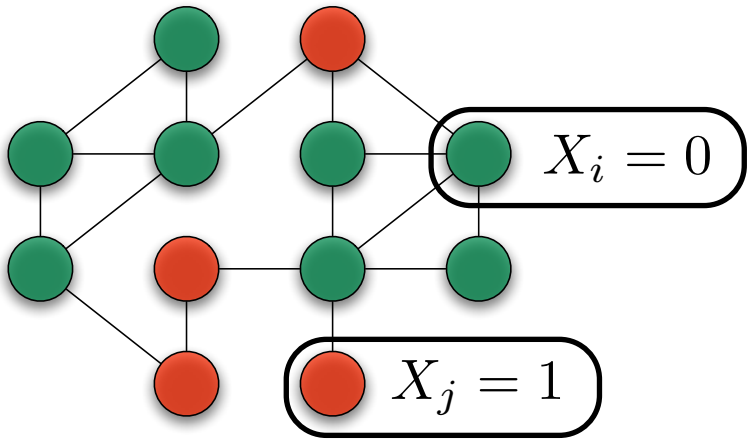
Let’s consider the cause of bias theoretically...

Statistical tests use
the *mean* and *variance*
of the *average error*

$$E_k = \frac{1}{n} \sum_{i=1}^{n} X_i$$



What is the mean and variance of E_k?



$$E(E_k) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$
$$= E(X)$$
$$= p \quad \textbf{Error rate}$$

Assuming the Xs are independent *Bernoulli(p)* random variables

$$Var(E_k) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$
$$= \frac{1}{n} Var(X)$$
$$= \frac{1}{n} p(1-p)$$

What if the errors are not independent?

Theorem 1:
Correlated errors increases variance

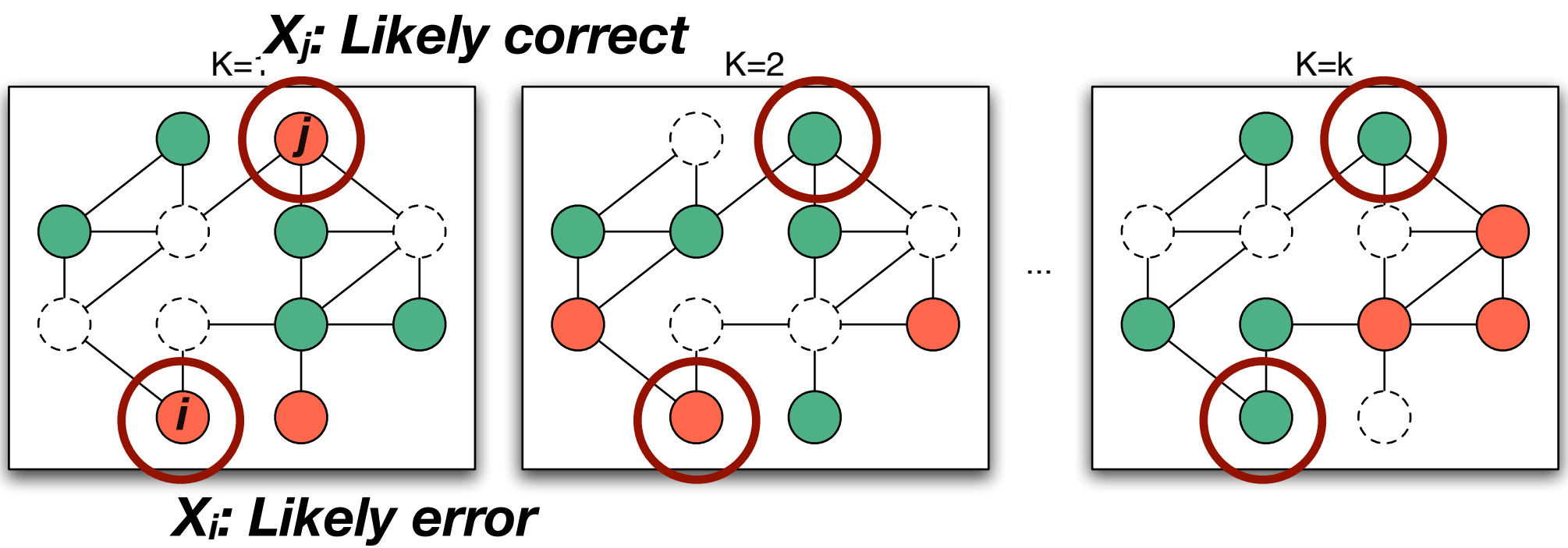
$$
\begin{aligned}
Var_{corr}(E_k) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} Var(X_i) + \sum_{i=1}^{n}\sum_{j\neq i}^{n} Cov(X_i, X_j)\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} p(1-p) + \sum_{i=1}^{n}\sum_{j\neq i}^{n} \frac{|E|}{n(n-1)}\rho \cdot p(1-p)\right) \\
&= \frac{1}{n^2}\left(n \cdot p(1-p) + |E|\rho \cdot p(1-p)\right) \\
&= \frac{1}{n} p(1-p)\left[1 + \rho\frac{|E|}{n}\right]
\end{aligned}
$$

Assuming linked pairs have correlation ρ

Additional variance due to correlation

How do overlapping samples affect variance?

Classification errors vary across test sets



Let the "likely error" rvs be distributed as Bernoulli(q)
Let the "likely correct" rvs be distributed as Bernoulli($\frac{p}{(1-p)}(1-q)$)

Q: represents variability of predictions across samples

Note: expected error rate is still p

Theorem 2:
Repeated sampling decreases variance

- If we have a graph with ***m*** nodes and sample repeatedly test sets of size ***n*** nodes

- If there are ***pm*** nodes with $\text{Bernoulli}(q)$ error and (***1-p)m*** nodes with $\text{Bernoulli}(\frac{p}{(1-p)}(1-q))$ error, then the variance of E_k is:

$$Var_{rs}(E_k) = \frac{1}{n}p(1-p)\left[1 - \frac{(n-1)}{(m-1)}\left(\frac{q-p}{1-p}\right)^2\right]$$

Variance underestimation increases with q

Theorem 3:

Repeated sampling + error correlation increases variance underestimation

- With repeated random sampling and error correlation... variance will be ***underestimated*** by:

$$\Delta = \frac{1}{n}p(1-p)\left[\frac{(n-1)}{(m-1)}\left(\frac{q-p}{1-p}\right)^2\right] + \frac{1}{n}p(1-p)\left[\rho\frac{|E|}{n}\left[1-\frac{1}{(m-1)}\left(\frac{1-q}{1-p}\right)\left[pmq-q+2mc\sqrt{pq}+mc^2-\frac{c^2}{(1-p)}\right]\right]\right]$$

Effect due to overlap

IID variance

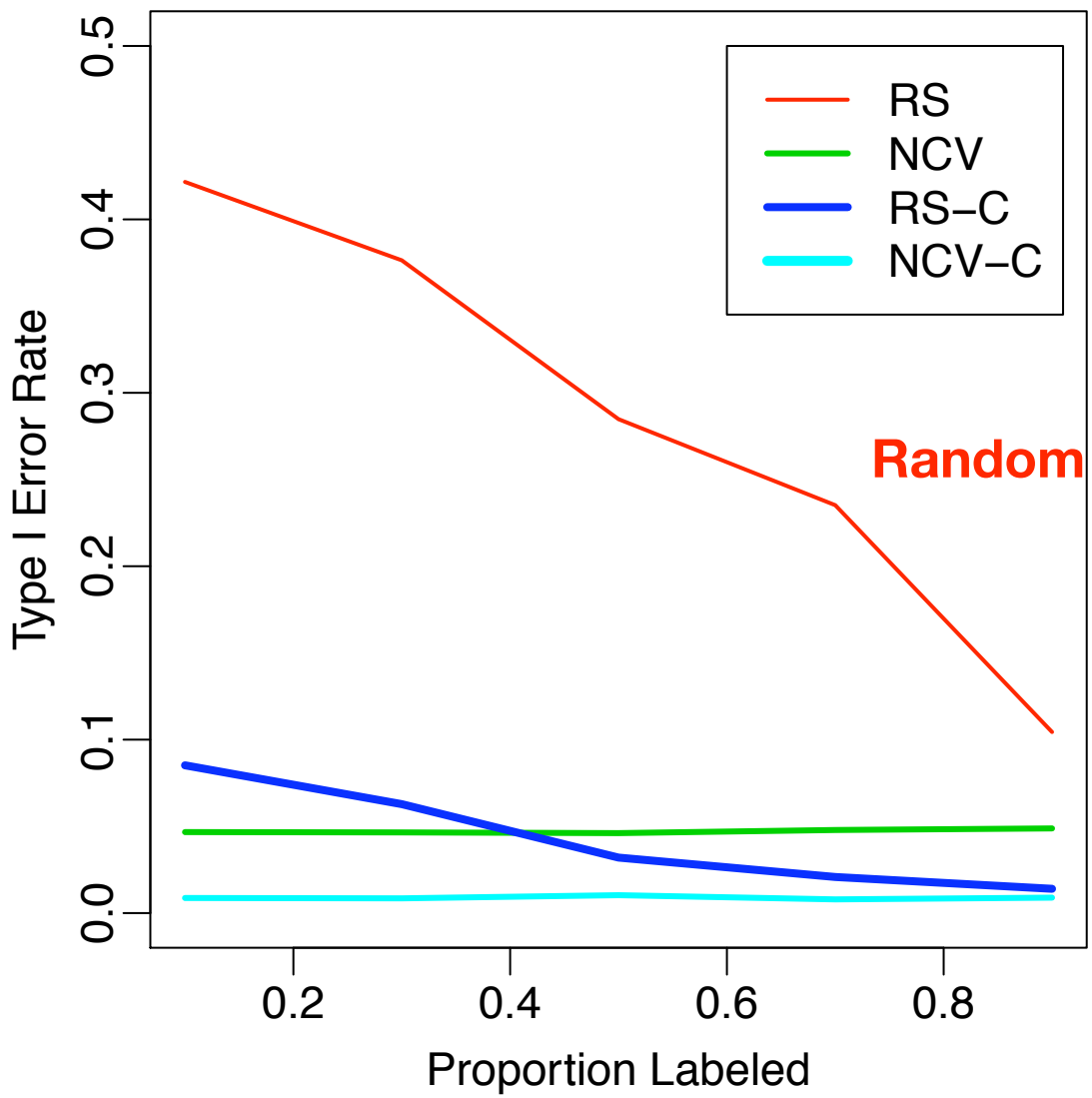
Extra effect due to overlap + correlation

- This increases the probability of Type 1 error by decreasing the critical value (t_α) used in the t-test

Analytical correction for bias *(Neville et al. 2010)*

- Calculate observed error variance
- Estimate ρ, ***p*** and ***q*** from samples
- Calculate underestimation effect due to resampling test sets of size ***n*** from network of size ***m***
- Add to observed variance to adjust for effect

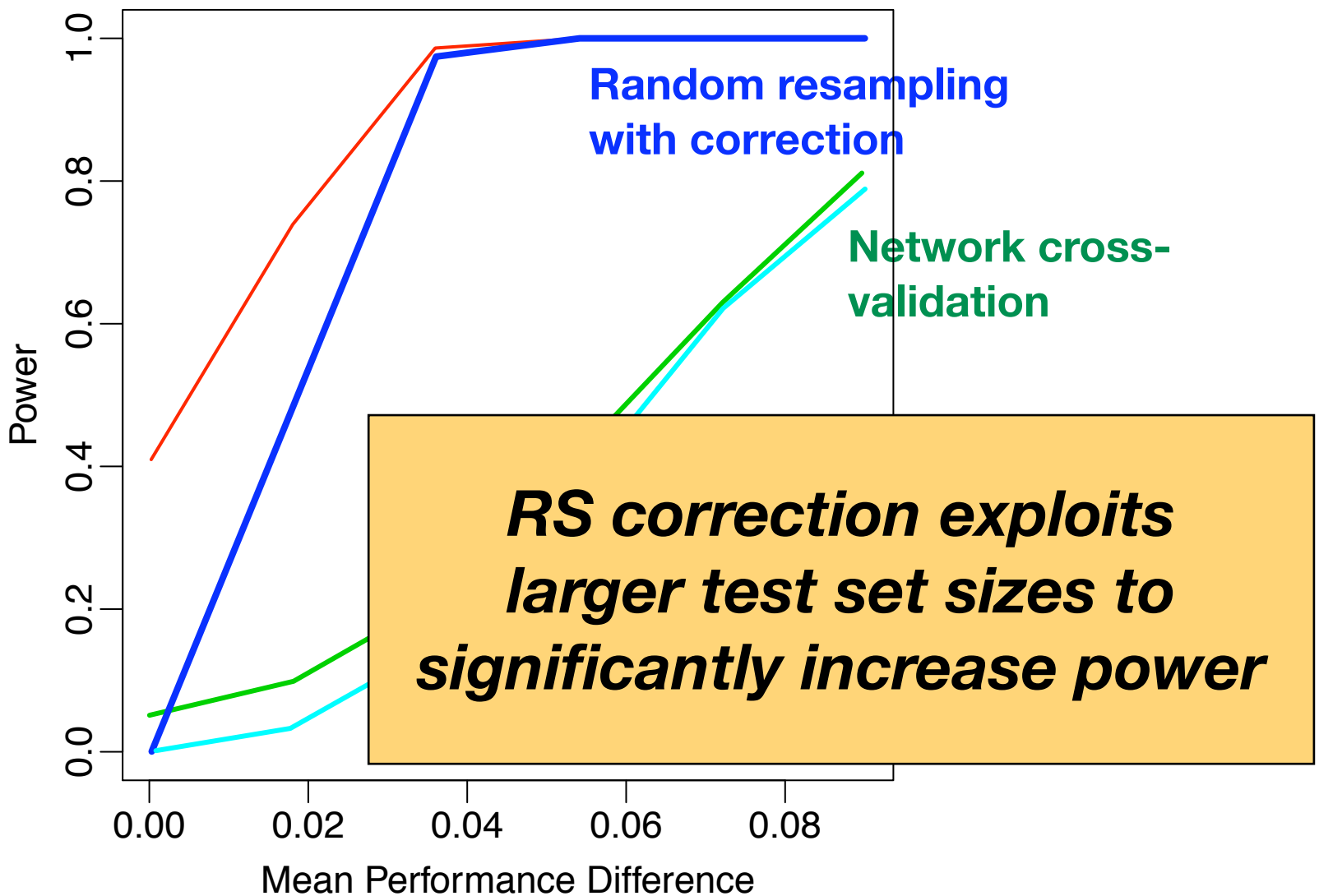
Analytical correction reduces Type I error



Random resampling

Random resampling with correction

Analytical correction increases statistical power



Our findings

- We show that commonly used statistical tests can result in **unacceptably high levels of Type I error** for network classifiers
  - This means that many algorithm differences will be judged incorrectly as significant when in fact performance is equivalent
- Solutions
  - **Network cross-validation**: Low Type I error, but decreased power
  - **Analytical correction**: Low Type I error, increased power
- Supported by broad set of empirical experiments
  - Synthetic data, simulated classifiers
  - Synthetic data, real classifiers
  - Real data, real classifiers

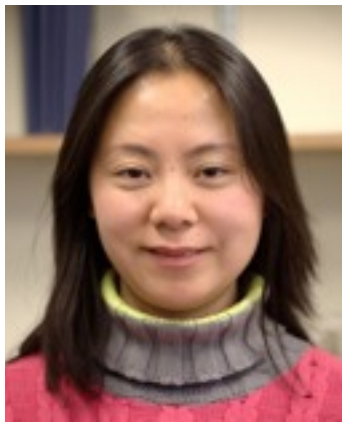
Applicability of results

- The bias will also affect:
  - More complex relational models -- since any relational model that attempts to exploit relational autocorrelation is likely to produce correlated errors
  - Across-network tasks -- if evaluation is on partially-labeled networks
  - Other forms of hypothesis testing in graphs (standard error may be underestimated)
- The extent of the bias will depend on:
  - Level of error correlation and amount of overlap between samples

Think carefully evaluation methodology for graph mining and classification algorithms...

in cases where the data consists of a <u>single</u> <u>network</u> with <u>heterogeneous</u> structure and <u>dependencies</u> among nodes-- naive application of conventional methods <u>can lead to incorrect conclusions</u>

Thanks to...



Tao Wang

Brian Gallagher

Tina Eliassi-Rad

To hear about other NLD research, talk to...

Nesreen Ahmed

Ryan Rossi

Joel Pfeiffer

Questions?

neville@cs.purdue.edu