Collective Opinion Spam Detection

Review Networks  Metadata

Shebuti Rayana

Leman Akoglu*

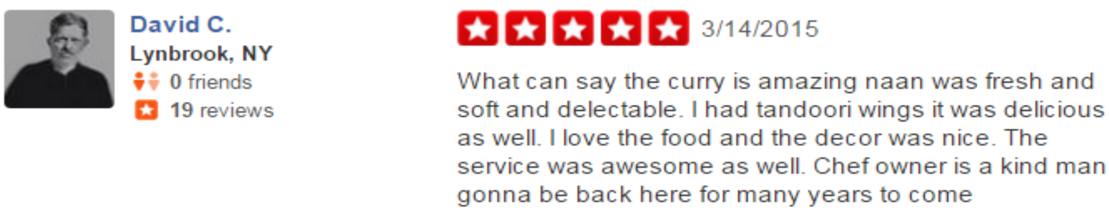
Stony Brook University

Computer Science

Introduction

- How do consumers learn about product quality?
  - Advertisements

  - Consumer review websites (Yelp, TripAdvisor etc.)



- Impact of consumer reviews on sales?

+1 star-rating increases revenue by 5-9%

Harvard Study by M. Luca
Reviews, Reputation, andRevenue: The Case ofYelp.com

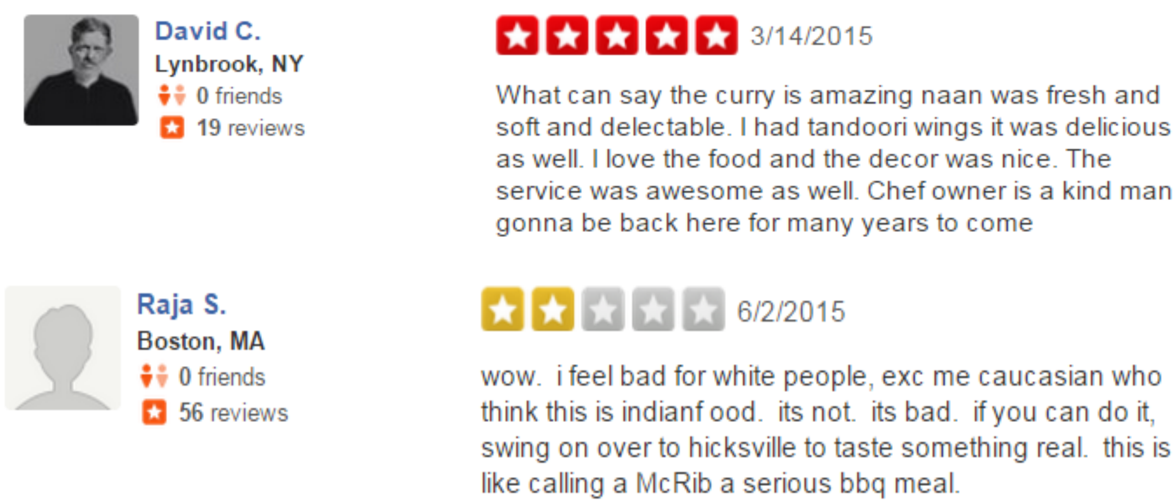
Opinion Spam

Stony Brook University Computer Science

- Paid/Biased reviewers write fake reviews
  - unjustly promote / demote products or businesses

Problem

?

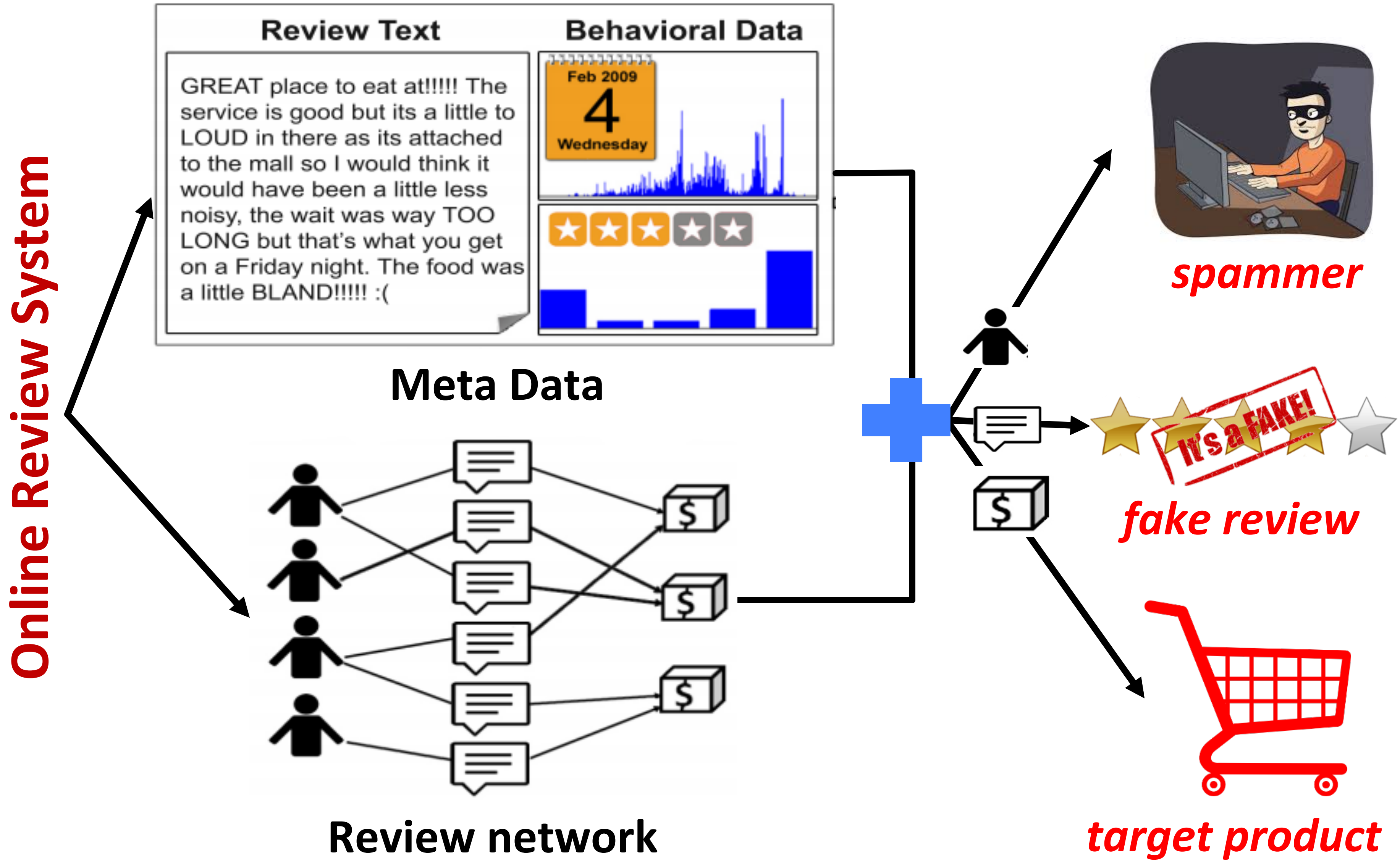


yelp

Humans only slightly better than chance

Finding Deceptive Opinion Spam by Any Stretch of the Imagination Ott et al. 2011

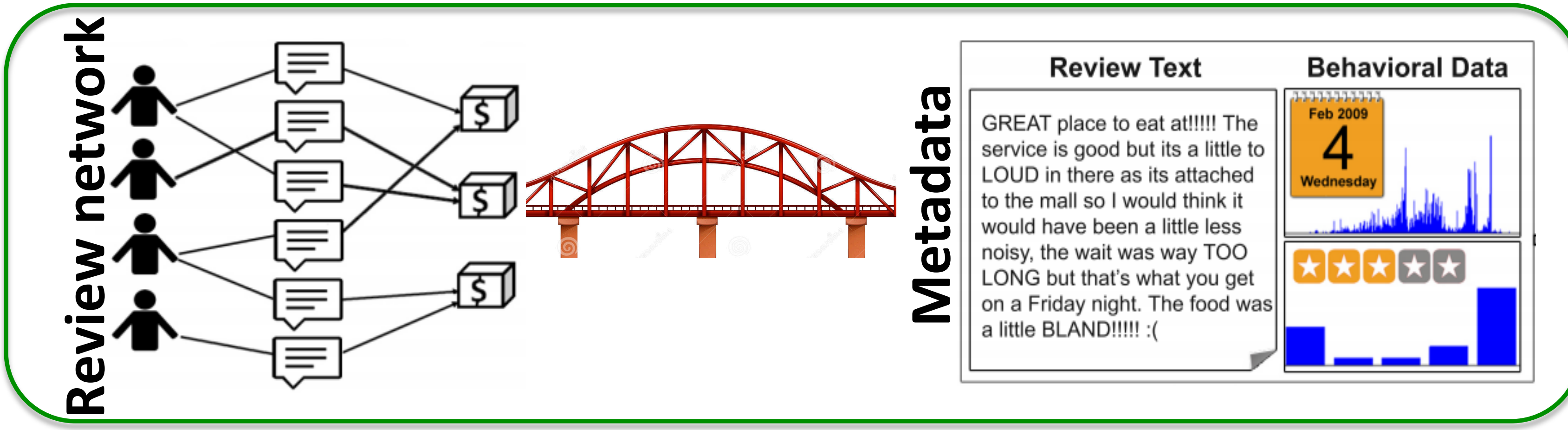
Goal: a “collective” approach



Overview

Main contributions:

- **SpEagle** : a **Collective** approach to opinion spam



  - is unsupervised
  - can easily leverage labels (**SpEagle⁺**)
  - improves detection performance
- Computationally light version : SpLite
  - significant speed-up

Related Work

Stony Brook University Computer Science

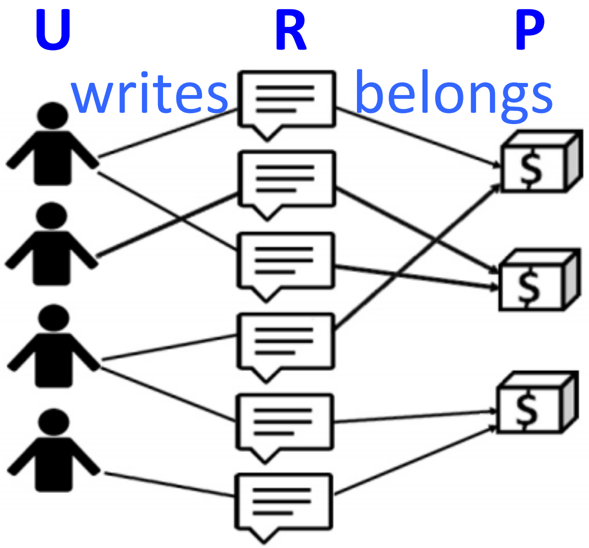
	Review Network	Review Text	Review Behavior	Supervision
Ott'2011		✓		supervised
Mukherjee' 2013		✓	✓	supervised
Jindal'2008			✓	supervised
Co-training [Li'2011]			✓	semi-supervised
Wang'2011	✓		✓	unsupervised
FraudEagle	✓			unsupervised
SpEagle	✓	✓	✓	unsupervised
SpEagle$^+$	✓	✓	✓	semi-supervised

Opinion Spam Detection: Problem

Stony Brook University Computer Science

A **network classification** problem

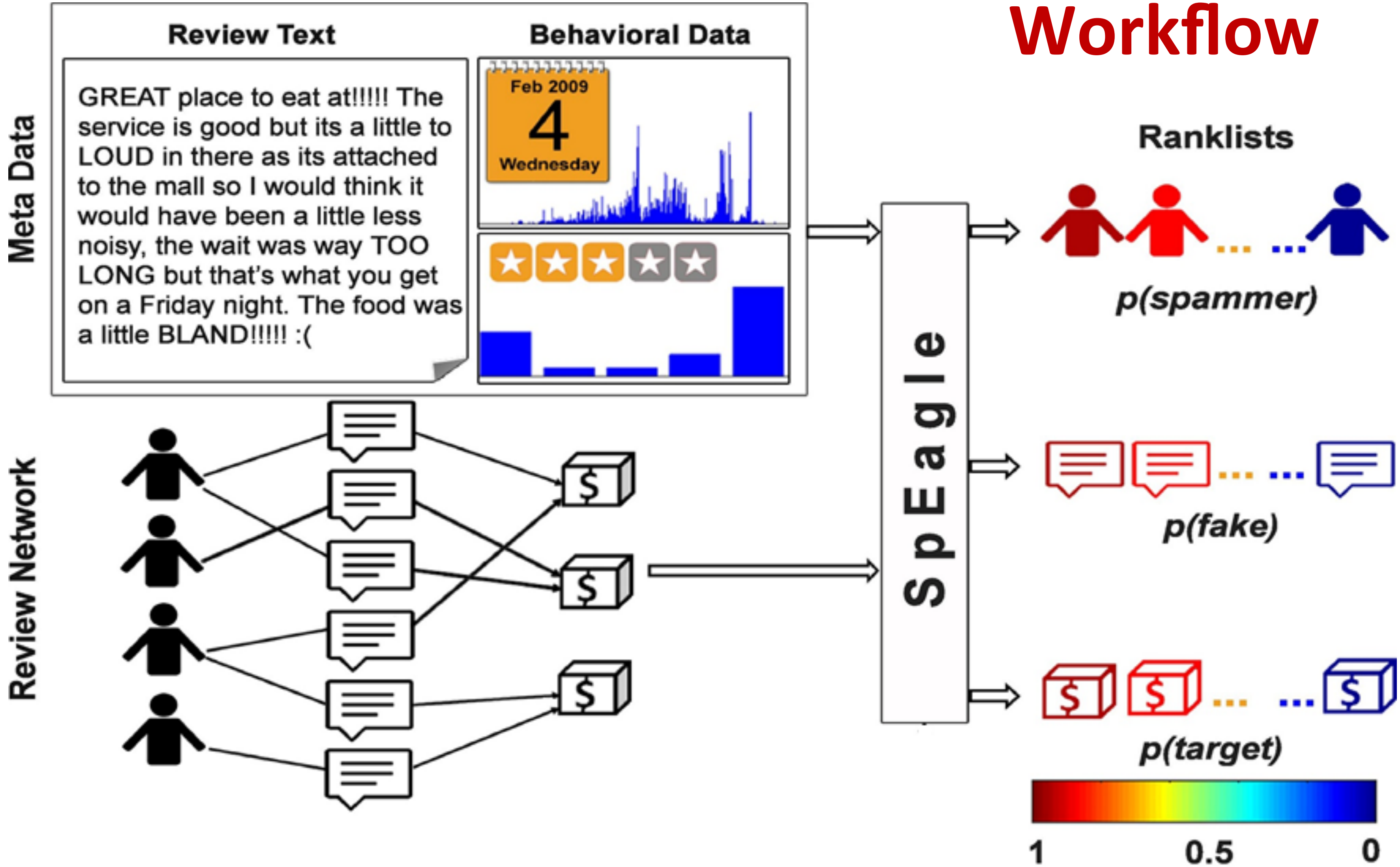
- **Given**
  - **U**ser-**R**eview-**P**roduct network (tri-partite)
  - Features extracted from metadata (i.e. text, behavior)
    - – for users, reviews, and products
- **Classify** network objects into type-specific classes
  - Users ('benign' vs. 'spammer')
  - Products ('non-target' vs. 'target')
  - Reviews ('genuine' vs. 'fake')



Proposed Approach: SpEagle

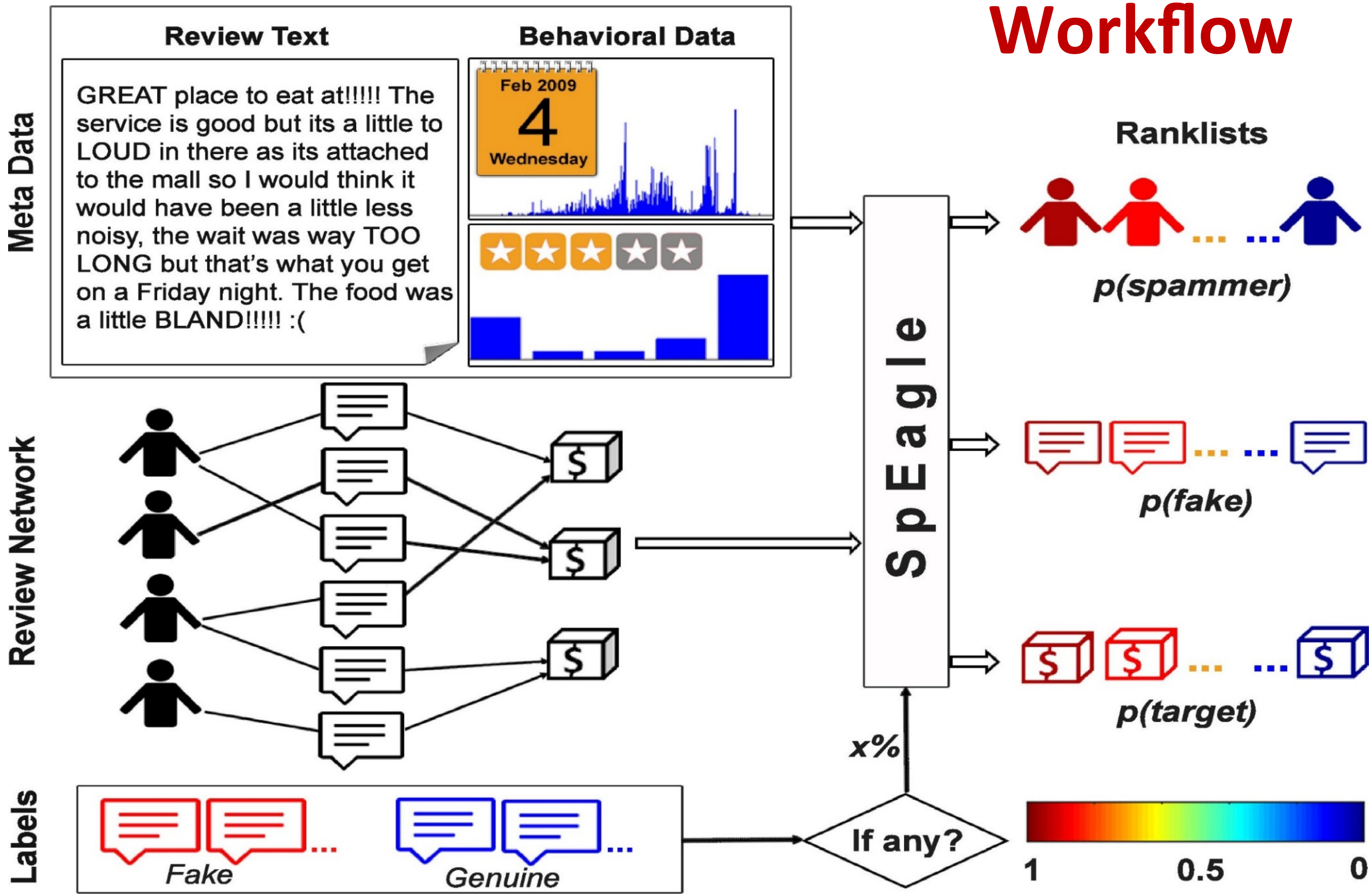
Stony Brook University
Computer Science

Workflow



Proposed Approach: SpEagle+

Workflow



SpEagle

Stony Brook University
Computer Science

- A collective classification approach (unsupervised)
  - Objective function utilizes pairwise Markov Random Fields

edge type

$$\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{Y_i \in \mathcal{Y}} \boxed{\phi_i(y_i)} \prod_{(Y_i, Y_j) \in E} \psi_{ij}^{t}(y_i, y_j)$$

Node labels as random variables

edge potential (label-label)

$$\boxed{\phi_i(y_i) = \psi_i(y_i) \prod_{(Y_i, X_j) \in E} \psi_{ij}^{t}(y_i)}$$

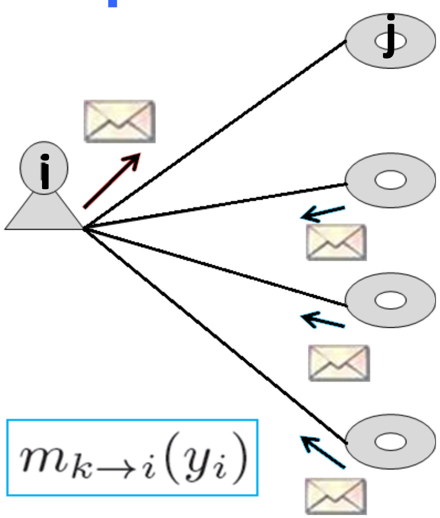
prior belief

edge potential (label-observed label)

SpEagle

- A collective classification approach (unsupervised)
  - Objective function utilizes pairwise Markov Random Fields
    - Inference problem (NP-hard)
- Loopy Belief Propagation (LBP)

1) Repeat for each node:



prior

edge type

edge potential

$$m_{i\rightarrow j}(y_j) = \alpha \sum_{y_i \in \mathcal{L}_{T_i}} \left(\boxed{\phi_i(y_i)} \boxed{\psi^{t}_{ij}(y_i, y_j)} \prod_{Y_k \in \mathcal{Y}_{\mathcal{N}_i} \setminus Y_j} m_{k\rightarrow i}(y_i) \right)$$

2) At convergence: $b_i(y_i) = \beta \ \phi_i(y_i) \prod_{Y_j \in \mathcal{Y}_{\mathcal{N}_i}} m_{j\rightarrow i}(y_i)$

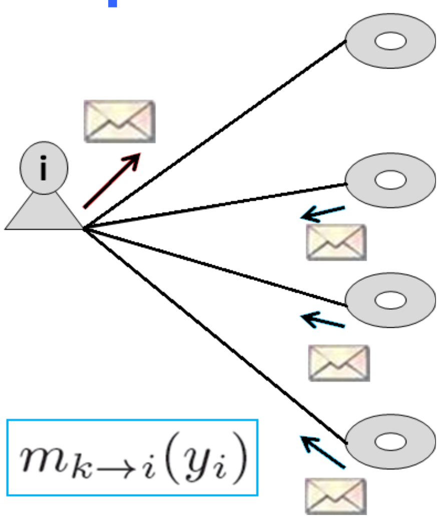
SpEagle

- A collective classification approach (unsupervised)
  - Objective function utilizes pairwise Markov Random Fields
    - Inference problem (NP-hard)
- Loopy Belief Propagation (LBP)

prior

edge type

1) Repeat for each node:



$$m_{i\to j}(y_j) = \alpha \sum_{y_i \in \mathcal{L}_{T_i}} \left(\boxed{\phi_i(y_i)}\, \boxed{\psi^{t}_{ij}(y_i, y_j)} \prod_{Y_k \in \mathcal{Y}_{\mathcal{N}_i} \setminus Y_j} m_{k\to i}(y_i) \right)$$

edge potential

2) At convergence: $b_i(y_i) = \beta\ \phi_i(y_i) \prod_{Y_j \in \mathcal{Y}_{\mathcal{N}_i}} m_{j\to i}(y_i)$

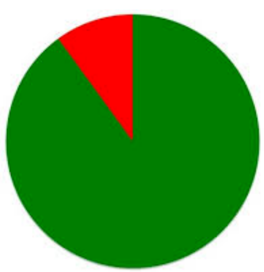
Priors

Metadata

Features

Spam Scores

Priors



Users: 'benign' 'spammer'

Products: 'non-target' 'target'

Reviews: 'genuine' 'fake'

Feature Extraction from Metadata

Stony Brook University
Computer Science

	User Features	Product Features	Review Features
Behavioral	• maximum #reviews/day • ratio of +ve/-ve reviews • avg/weighted rating deviation • rating deviation entropy • temporal gaps entropy • burstiness of reviews	• maximum #reviews/day • ratio of +ve/-ve reviews • avg/weighted rating deviation • rating deviation entropy • temporal gaps entropy	• rank order of reviews • absolute/thresholded rating deviation • extremity of rating • early time frame • singleton review
Text	• review length (#words) • avg content similarity • max content similarity	• review length • average content similarity • maximum content similarity	• ratio subjective/objective • description length • ratio of exclamation sent. • freq. of similar reviews • % capital letters • review length • ratio 1^{st} person pronoun

Feature Extraction from Metadata

Stony Brook University
Computer Science

	User Features	Product Features	Review Features
Behavioral	• maximum #reviews/day • ratio of +ve/-ve reviews • avg/weighted rating deviation • rating deviation entropy • temporal gaps entropy • burstiness of reviews	• maximum #reviews/day • ratio of +ve/-ve reviews • avg/weighted rating deviation • rating deviation entropy • temporal gaps entropy	• rank order of reviews • absolute/thresholded rating deviation • extremity of rating • early time frame • singleton review
Text	• review length (#words) • avg content similarity • max content similarity	• review length • avg content similarity • max content similarity	• ratio subjective/objective • description length • ratio of exclamation sent. • freq. of similar reviews • % capital letters • review length • ratio 1st person pronoun

Feature Extraction from Metadata

Stony Brook University Computer Science

	User Features	Product Features	Review Features
Behavioral	• maximum #reviews/day • ratio of +ve/-ve reviews • avg/ **Weighted Rating Deviation** • rating deviation entropy • temporal gaps entropy • burstiness of reviews	• maximum #reviews/day • ratio of +ve/-ve reviews • avg/ **Weighted Rating Deviation** • rating deviation entropy • temporal gaps entropy	• rank order of reviews • absolute/thresholded rating deviation • extremity of rating • early time frame • singleton review
Text	• review length (#words) • avg content similarity • max content similarity	• review length • avg content similarity • max content similarity	• ratio subjective/objective • **Description Length** • ratio of exclamation sent. • freq. of similar reviews • % capital letters • review length • ratio 1st person pronoun

$$\boxed{\frac{\sum_{e_{ij} \in E_{i*}} |d_{ij}| w_{ij}}{\sum_{e_{ij} \in E_{i*}} w_{ij}}}$$

$$d_{ij} = r_{ij} - avg_{e \in E_{*j}} r(e)$$

$$w_{ij} = \frac{1}{(t_{ij})^{\alpha}}$$

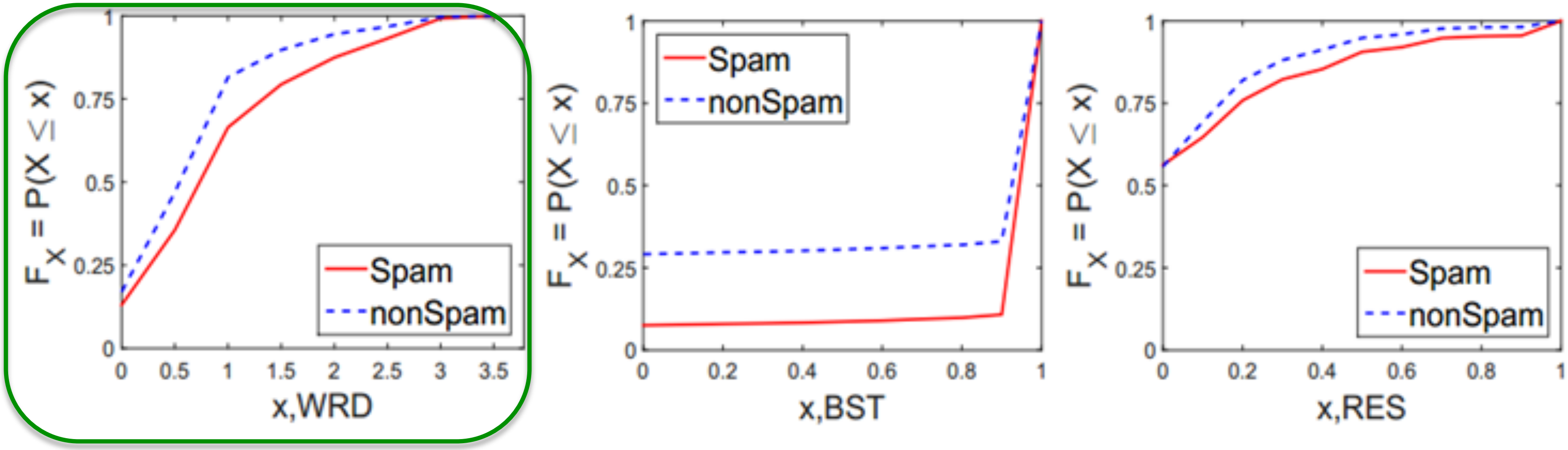
temporal order

$$\boxed{\sum_{w} -log(freq(w))}$$

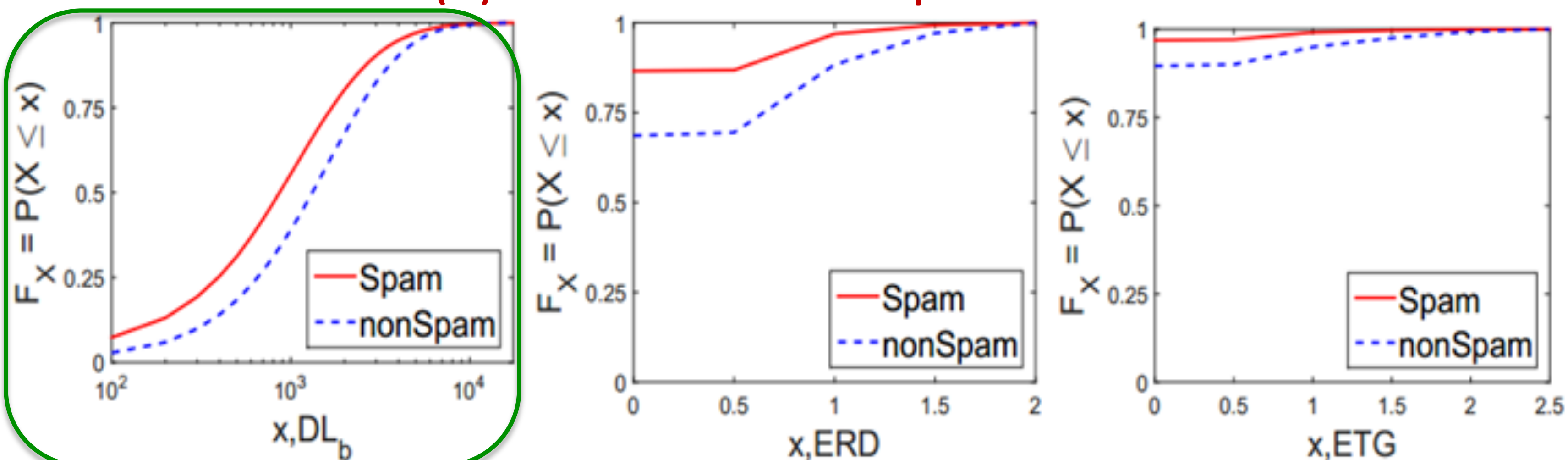
words in review

Feature Analysis

(H)igher more suspicious



(L)ower more suspicious



Spam Score & Prior Computation

Stony Brook University
Computer Science

Q: How to handle features with different scales?

A: Cumulative distribution:

- For each feature l , $1 \leq l \leq F$ and its corresponding value x_{li} for node i

$$f(x_{li}) = \begin{cases} 1 - P(X_l \leq x_{li}), & \textit{if high is suspicious } (H) \\ P(X_l \leq x_{li}), & \textit{otherwise } (L) \end{cases}$$

- Combine F values for each node i:

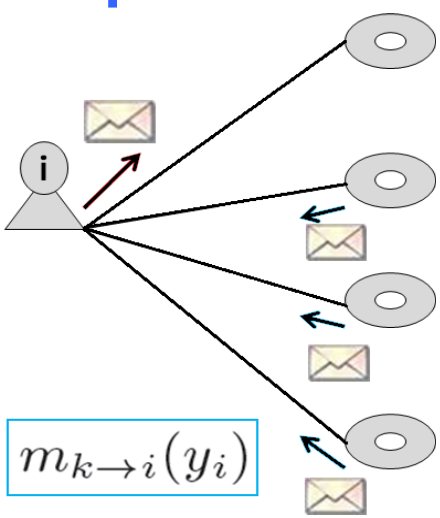
$$S_i = 1 - \sqrt{\frac{\sum_{l=1}^{F} f(x_{li})^2}{F}}$$

- Priors : $\phi_i \leftarrow \{1 - S_i, S_i\}$

SpEagle

- A collective classification approach (unsupervised)
  - Objective function utilizes pairwise Markov Random Fields
    - Inference problem (NP-hard)
- Loopy Belief Propagation (LBP)

1) Repeat for each node:

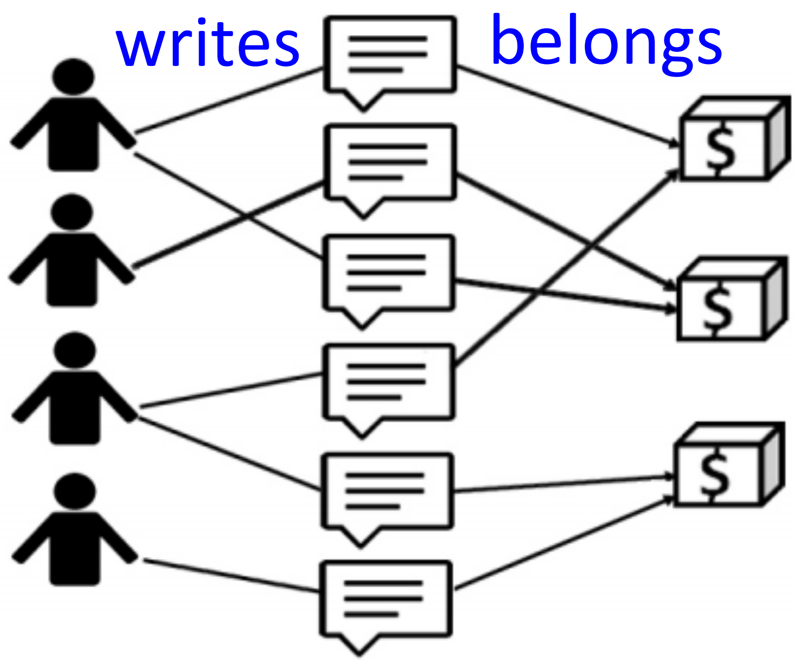


$$m_{i\rightarrow j}(y_j) = \alpha \sum_{y_i \in \mathcal{L}_{T_i}} \Bigg(\overbrace{\boxed{\phi_i(y_i)}}^{\text{prior}} \underbrace{\boxed{\psi_{ij}^{\overset{\text{edge type}}{t}}(y_i, y_j)}}_{\text{edge potential}} \prod_{Y_k \in \mathcal{Y}_{\mathcal{N}_i} \setminus Y_j} m_{k\rightarrow i}(y_i) \Bigg)$$

2) At convergence: $b_i(y_i) = \beta\ \phi_i(y_i) \prod_{Y_j \in \mathcal{Y}_{\mathcal{N}_i}} m_{j\rightarrow i}(y_i)$

Edge Potentials

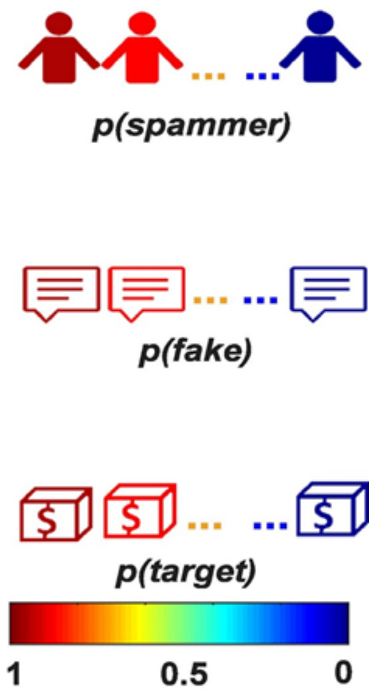
	User ($\psi^{t=\text{'write'}}$)		($\psi^{t=\text{'belong'}}$) **Product**	
Review	*benign*	*spammer*	*non-target*	*target*
genuine	1	0	$1-\epsilon$	ϵ
fake	0	1	ϵ	$1-\epsilon$



Classification

Beliefs as class probabilities:

- $Prob_i(spammer) = b_i(y_i : spammer)$
- $Prob_k(fake) = b_k(y_k : fake)$
- $Prob_j(target) = b_j(y_j : target)$

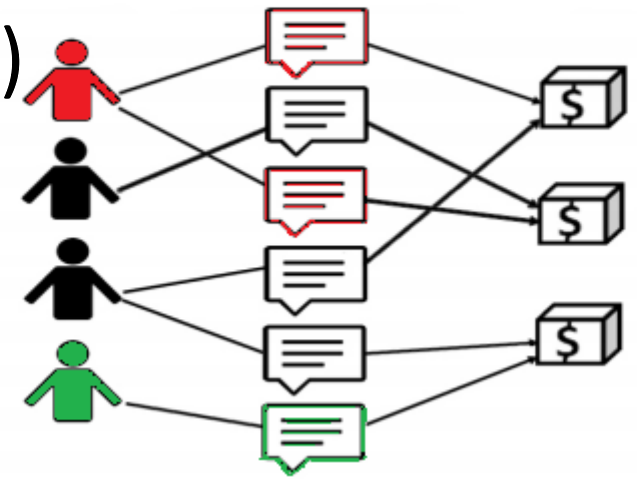


SpEagle^{+}: Leveraging Labels

Stony Brook University
Computer Science

- SpEagle can work semi-supervised
  - Can incorporate labels seamlessly
  - Can use user, review, and/or product labels

- For labeled nodes, priors are set to:
  - $\varphi \leftarrow \{\epsilon, 1-\epsilon\}$ for spam category (i.e., *fake*, *spammer*, or *target*)
  - $\varphi \leftarrow \{1-\epsilon, \epsilon\}$ for non-spam category



Data Sets

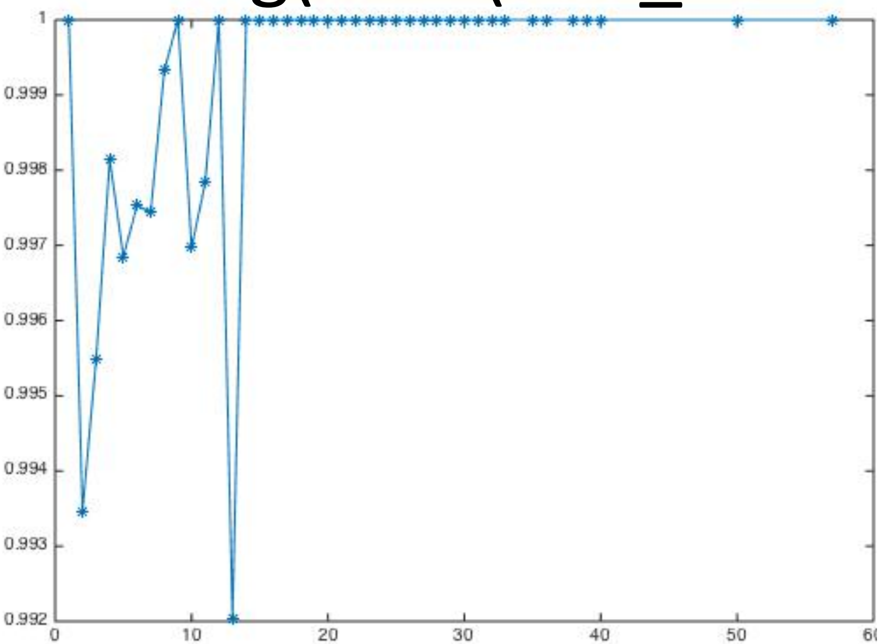
- 3 **Yelp** datasets[1]: recommended vs. non-recommended
  - **YelpChi** –hotel and restaurant reviews from Chicago
  - **YelpNYC** –restaurant reviews from New York City
  - **YelpZip** –restaurants reviews from zipcodes in NJ, VT, CT, PA

Dataset	#Reviews (filtered %)	#Users (spammer %)[2]	#Products (rest.&hotel)
YelpChi	67,395 (13.23%)	38,063 (20.33%)	201
YelpNYC	359,052 (10.27%)	160,225 (17.79%)	923
YelpZip	608,598 (13.22%)	260,277 (23.91%)	5,044

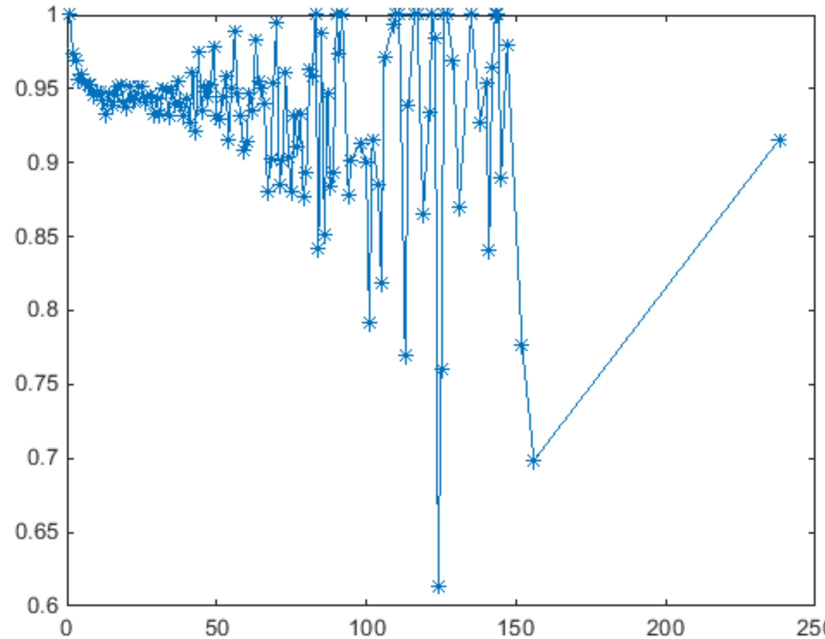
[1] Datasets are made available to the community
[2] A spammer has at least one filtered review

Labeling users with >0 filtered reviews as spammers

Avg(max(frac_filtered, frac_nonfiltered))



#reviews

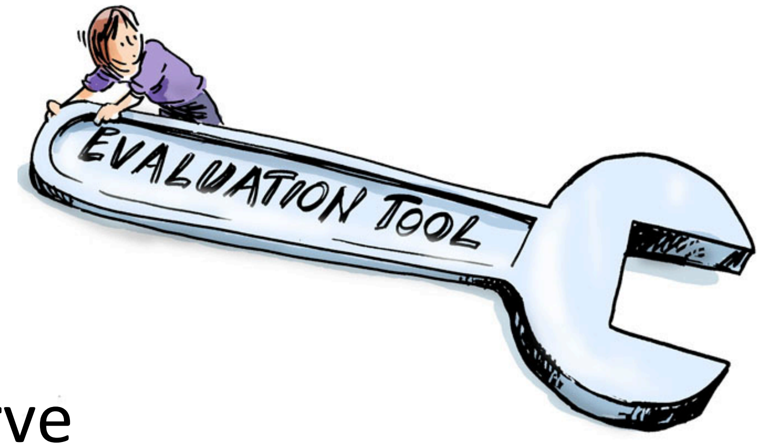


#reviews

Evaluation Metrics

Stony Brook University
Computer Science

- Area Under Curve (AUC)
  - for ROC curve (TPR vs. FPR)



- Average Precision (AP)
  - AUC for Precision-Recall curve
- Precision@k : ratio of spam in top *k*
- NDCG@k *:* weighted scoring which favors top items

$$NDCG@k = \frac{DCG@k}{IDCG@k} \text{ for } DCG@k = \sum_{i=1}^{k} \frac{2^{l_i} - 1}{\log_2(i+1)}$$

Experiment Results

- SpEagle superior to existing methods

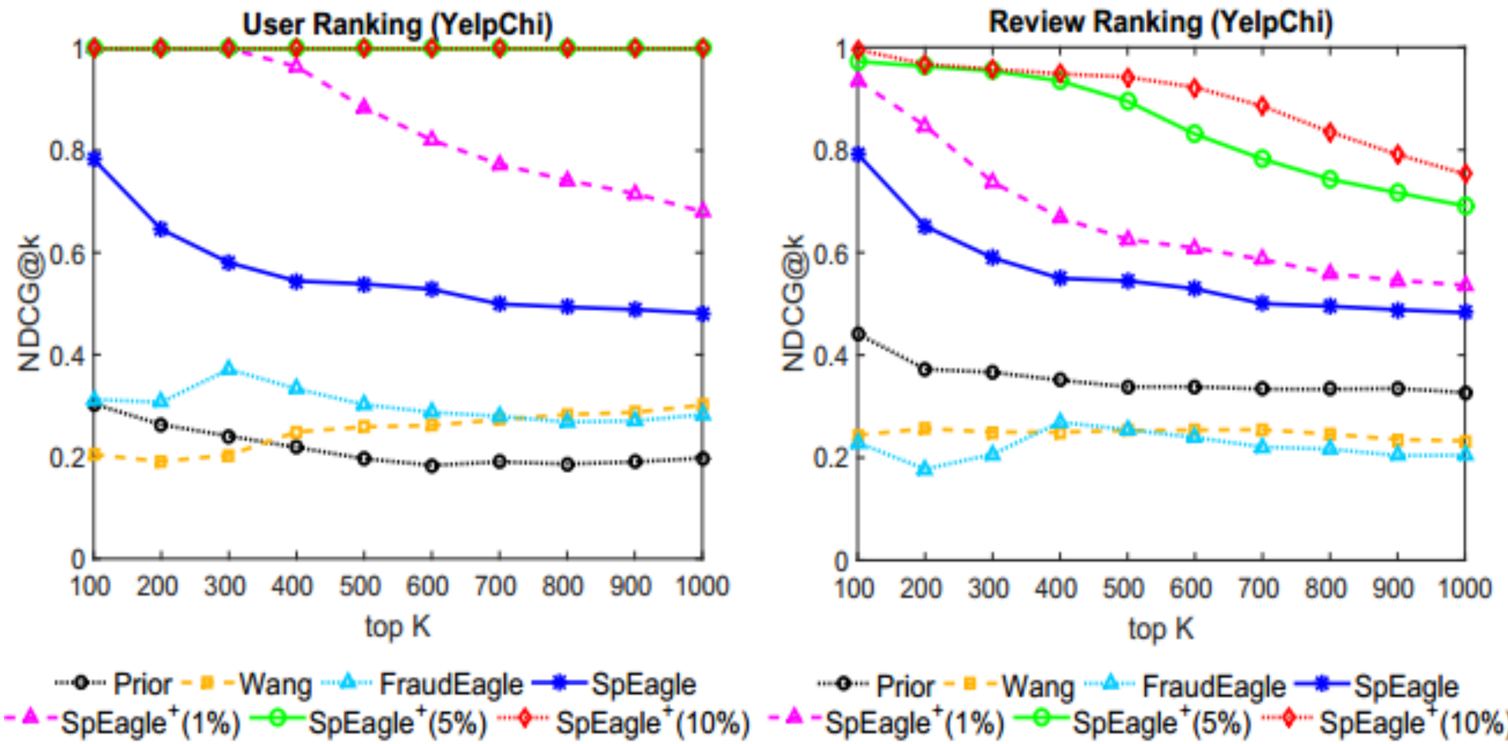
	User Ranking						Review Ranking					
	AP			AUC			AP			AUC		
	Y'Chi	Y'NYC	Y'Zip	Y'Chi	Y'NYC	Y'Zip	Y'Chi	Y'NYC	Y'Zip	Y'Chi	Y'NYC	Y'Zip
RANDOM	0.2024	0.1782	0.2392	0.5000	0.5000	0.5000	0.1327	0.1028	0.1321	0.5000	0.5000	0.5000
FRAUDEAGLE	0.2537	0.2233	0.3091	0.6124	0.6062	0.6175	0.1067	0.1122	0.1524	0.3735	0.5063	0.5326
WANG ET AL.	0.2659	0.2381	0.3306	0.6167	0.6207	0.6554	0.1518	0.1255	0.1803	0.5062	0.5415	0.5982
PRIOR	0.2157	0.1826	0.2550	0.5294	0.5081	0.5269	0.2241	0.1789	0.2352	0.6707	0.6705	0.6838
SPEAGLE	**0.3393**	**0.2680**	**0.3616**	**0.6905**	**0.6575**	**0.6710**	**0.3236**	**0.2460**	**0.3319**	**0.7887**	**0.7695**	**0.7942**
SPEAGLE^{+}(1%)	0.3967	0.3480	0.4245	0.7078	0.6828	0.6907	0.3352	0.2757	0.3545	0.7951	0.7829	0.8040
SPLITE^{+} (1%)	0.3777	0.3331	0.4218	0.6744	0.6542	0.6784	0.3124	0.2550	0.3448	0.7693	0.7631	0.7923

- Different priors: **U**ser & **R**eview priors most informative

	User Ranking						Review Ranking					
	AP			AUC			AP			AUC		
	Y'Chi	Y'NYC	Y'Zip	Y'Chi	Y'NYC	Y'Zip	Y'Chi	Y'NYC	Y'Zip	Y'Chi	Y'NYC	Y'Zip
RANDOM	0.2024	0.1782	0.2392	0.5000	0.5000	0.5000	0.1327	0.1028	0.1321	0.5000	0.5000	0.5000
SPEAGLE (U)	0.3197	0.2624	0.2808	0.6767	0.6483	0.6183	0.3043	0.2400	0.1427	0.7783	0.7629	0.5940
SPEAGLE (P)	0.1550	0.1357	0.1814	0.3905	0.3930	0.3801	0.0755	0.0640	0.0806	0.1643	0.2536	0.2277
SPEAGLE (R)	0.3226	0.2575	0.3449	0.6771	0.6477	0.6562	0.3098	0.2378	0.3180	0.7820	0.7656	0.7884
SPEAGLE (UR)	**0.3398**	**0.2680**	0.3615	**0.6905**	**0.6575**	0.6709	**0.3241**	**0.2460**	**0.3320**	**0.7887**	**0.7695**	**0.7942**
SPEAGLE (URP)	0.3393	**0.2680**	**0.3616**	**0.6905**	**0.6575**	**0.6710**	0.3236	**0.2460**	0.3319	**0.7887**	**0.7695**	**0.7942**

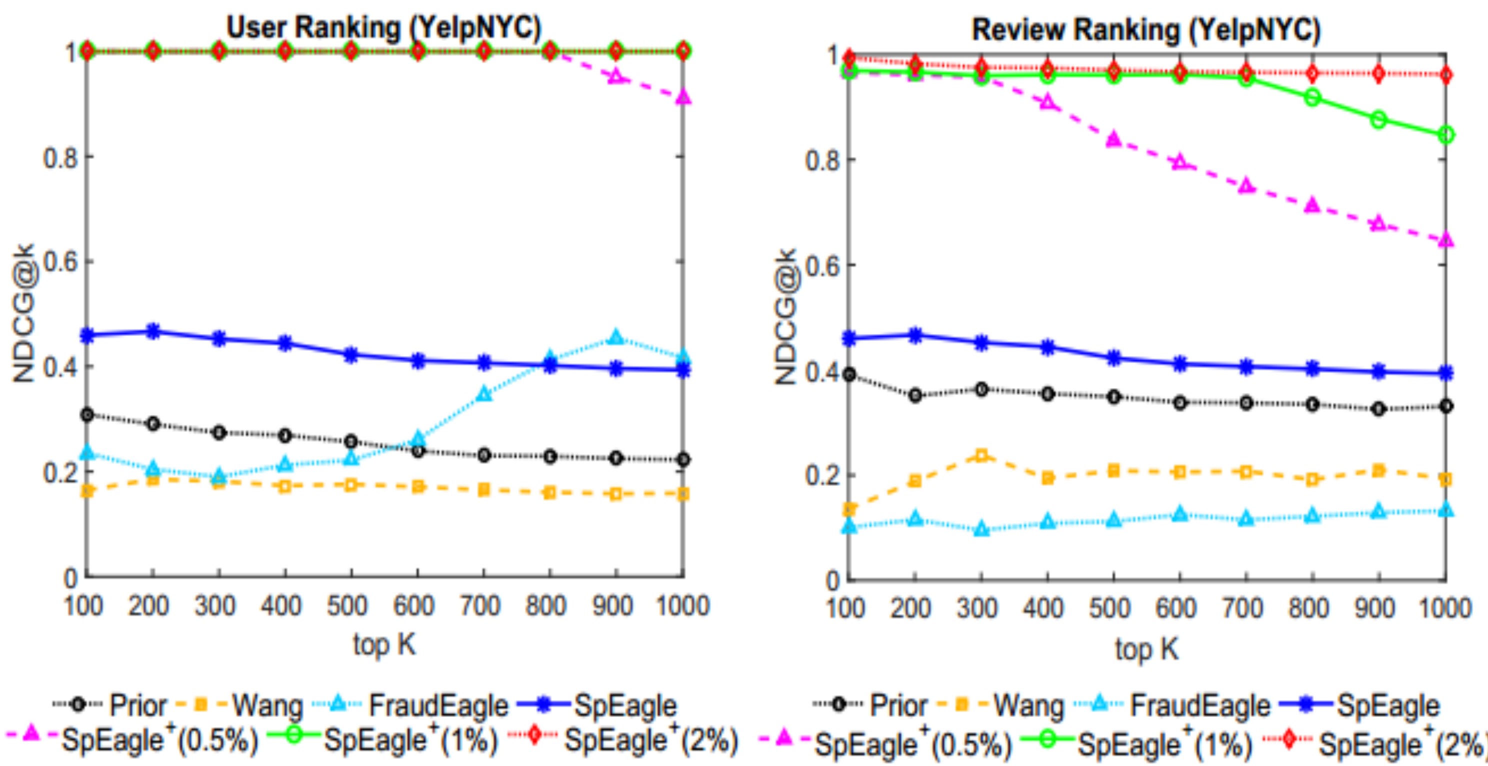
NDCG@k : YelpChi

- Labels improve performance significantly



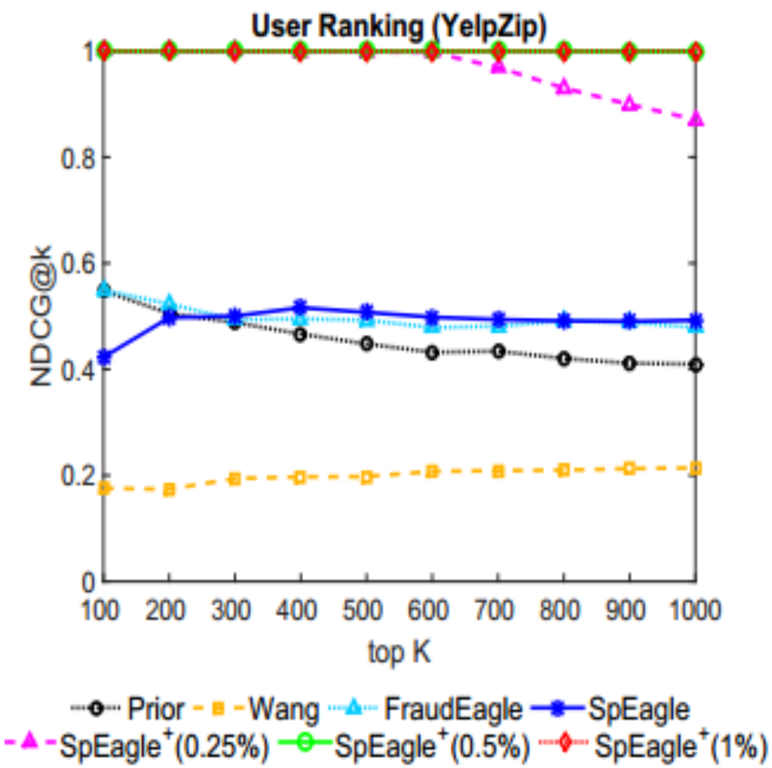
NDCG@k : YelpNYC

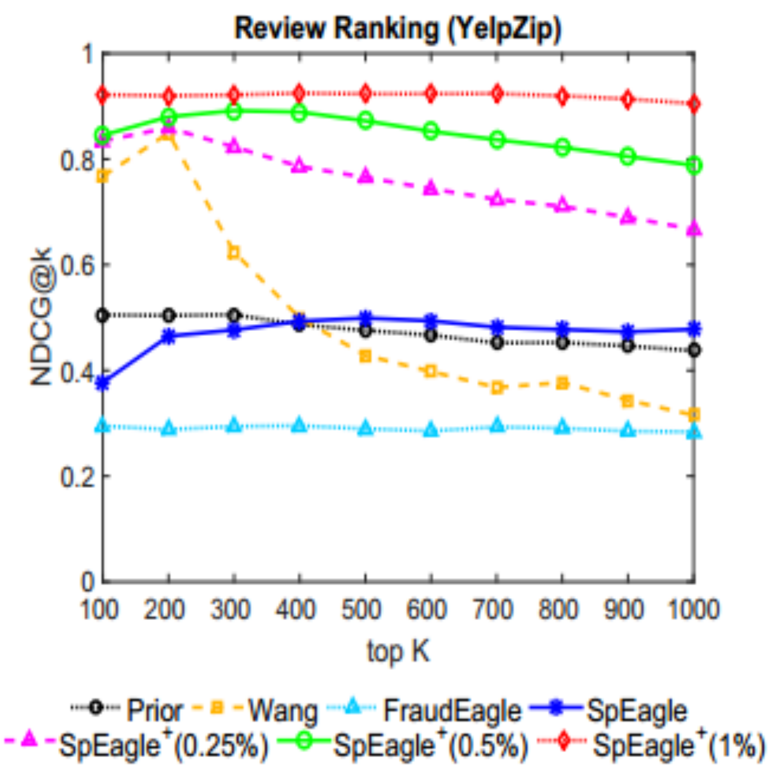
- Labels improve performance significantly



NDCG@k : YelpZip

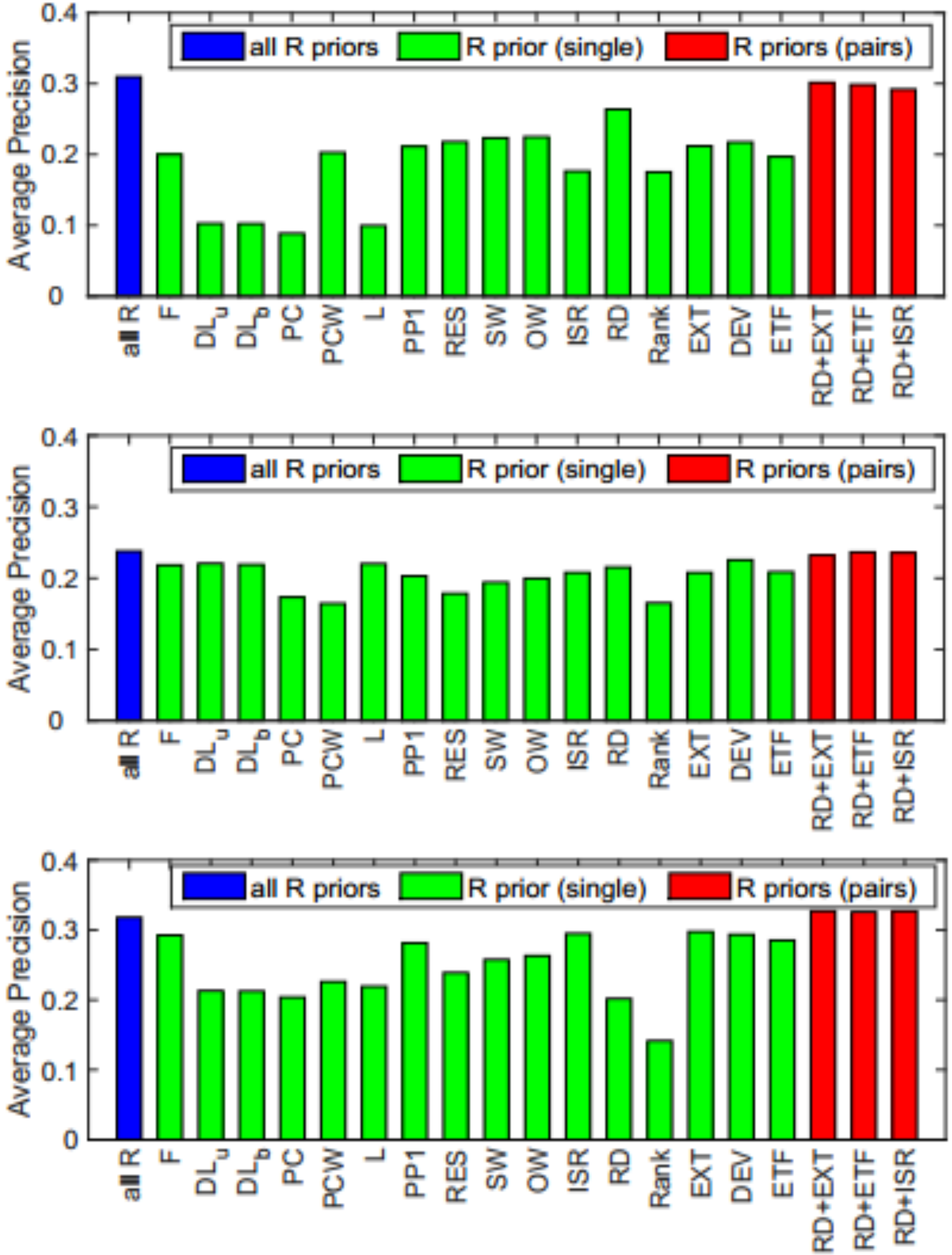
- Labels improve performance significantly





SpLite (SpLite$^+$)

Light version of **SpEagle**



Review features:

all

individual

pairs

SpLite: Experimental Results

Stony Brook University Computer Science

	User Ranking						Review Ranking					
	AP			AUC			AP			AUC		
	Y'Chi	Y'NYC	Y'Zip	Y'Chi	Y'NYC	Y'Zip	Y'Chi	Y'NYC	Y'Zip	Y'Chi	Y'NYC	Y'Zip
Random	0.2024	0.1782	0.2392	0.5000	0.5000	0.5000	0.1327	0.1028	0.1321	0.5000	0.5000	0.5000
FraudEagle	0.2537	0.2233	0.3091	0.6124	0.6062	0.6175	0.1067	0.1122	0.1524	0.3735	0.5063	0.5326
Wang et al.	0.2659	0.2381	0.3306	0.6167	0.6207	0.6554	0.1518	0.1255	0.1803	0.5062	0.5415	0.5982
Prior	0.2157	0.1826	0.2550	0.5294	0.5081	0.5269	0.2241	0.1789	0.2352	0.6707	0.6705	0.6838
SpEagle	**0.3393**	**0.2680**	**0.3616**	**0.6905**	**0.6575**	**0.6710**	**0.3236**	**0.2460**	**0.3319**	**0.7887**	**0.7695**	**0.7942**
SpEagle$^+$ (1%)	0.3967	0.3480	0.4245	0.7078	0.6828	0.6907	0.3352	0.2757	0.3545	0.7951	0.7829	0.8040
SpLite$^+$ (1%)	0.3777	0.3331	0.4218	0.6744	0.6542	0.6784	0.3124	0.2550	0.3448	0.7693	0.7631	0.7923

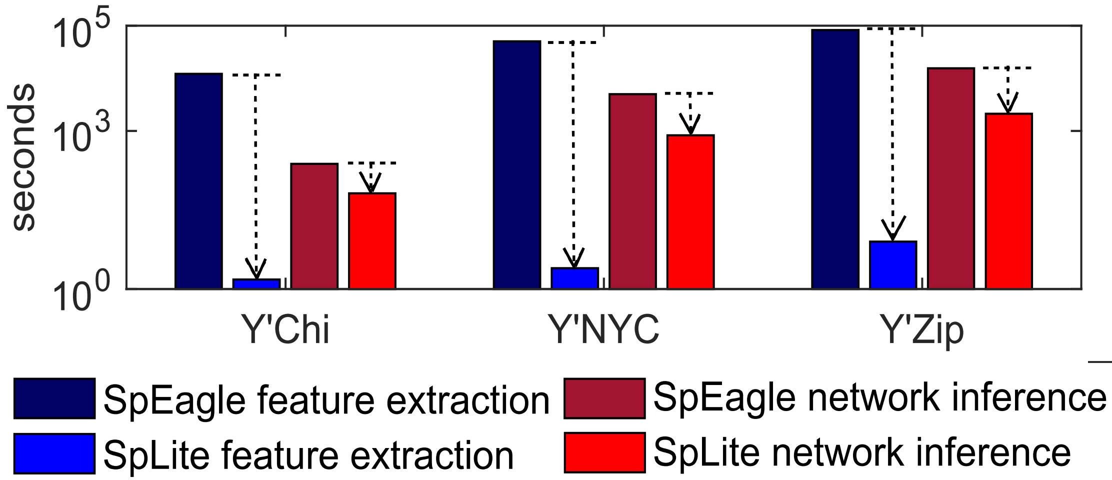
- SpEagle$^+$ vs SpLite$^+$ perform comparably

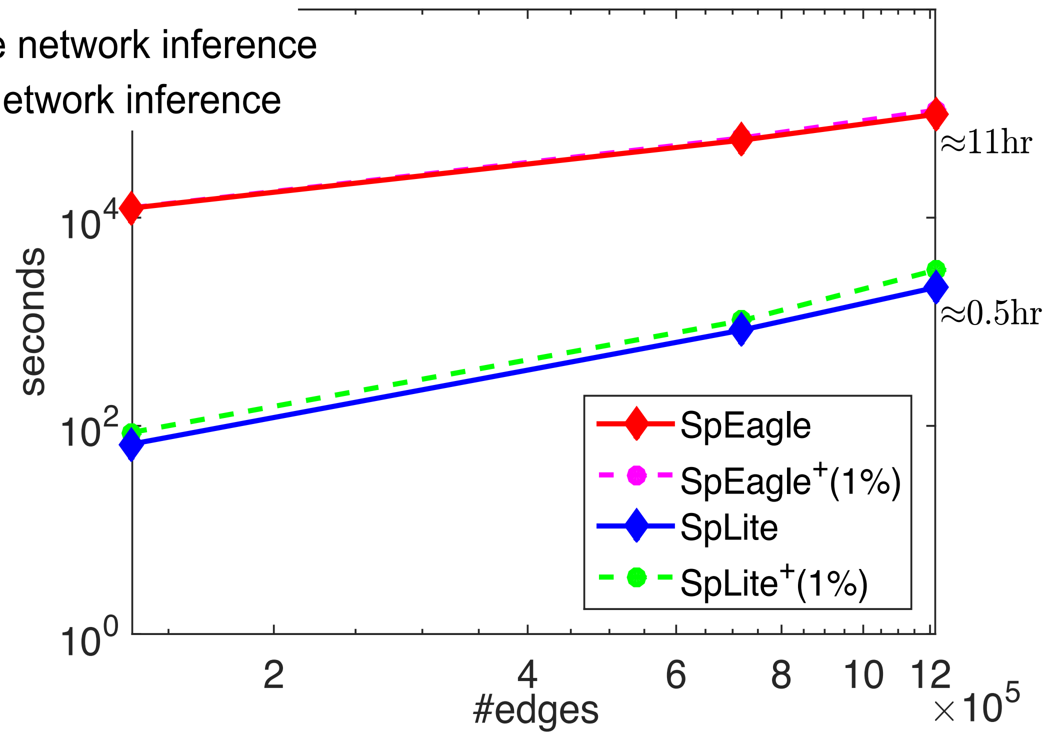
Table 8: *NDCG@k* **performance comparison of SpEagle vs. SpLite (with 1% supervision on all datasets)**

	User Ranking						Review Ranking					
	YelpChi		YelpNYC		YelpZip		YelpChi		YelpNYC		YelpZip	
k	Sp'le	SpLite	Sp'le	SpLite	Sp'le	SpLite	Sp'le	SpLite	Sp'le	SpLite	Sp'le	SpLite
100	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9354	0.9334	0.9694	0.9651	0.9219	0.9377
200	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8469	0.8007	0.9665	0.9595	0.9200	0.9379
300	1.0000	0.9995	1.0000	1.0000	0.9997	1.0000	0.7373	0.6986	0.9597	0.9584	0.9216	0.9377
400	0.9645	0.9589	1.0000	1.0000	0.9998	1.0000	0.6682	0.6397	0.9615	0.9571	0.9248	0.9360
500	0.8841	0.8677	1.0000	1.0000	0.9998	1.0000	0.6255	0.6103	0.9610	0.9529	0.9234	0.9276
600	0.8205	0.8107	1.0000	1.0000	0.9998	1.0000	0.6089	0.5740	0.9620	0.9432	0.9236	0.9121
700	0.7731	0.7650	1.0000	1.0000	0.9999	1.0000	0.5864	0.5556	0.9552	0.8925	0.9240	0.9021
800	0.7416	0.7279	1.0000	1.0000	0.9999	1.0000	0.5587	0.5317	0.9179	0.8351	0.9199	0.8977
900	0.7157	0.6980	1.0000	1.0000	0.9999	1.0000	0.5458	0.5279	0.8775	0.7923	0.9138	0.8899
1000	0.6803	0.6670	1.0000	1.0000	0.9999	1.0000	0.5361	0.5218	0.8463	0.7577	0.9052	0.8810

Running Time & Scalability

- SpLite+ is orders of magnitude faster than **SpEagle+**



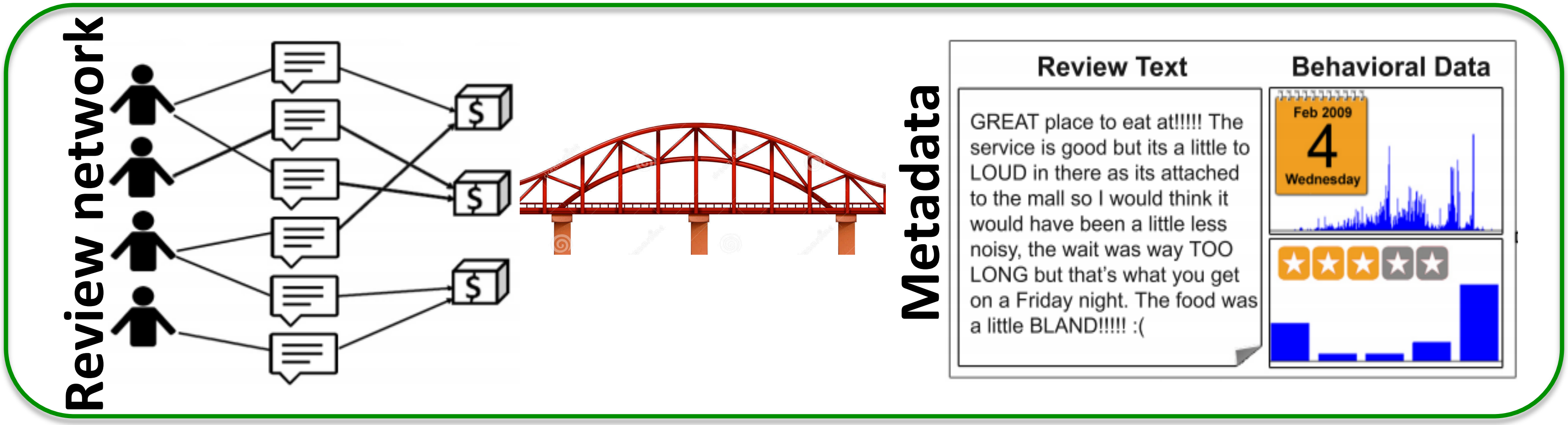


Summary

Stony Brook University
Computer Science

Main contributions:

- **SpEagle** : a **Collective** approach to opinion spam



  - is unsupervised
  - can easily leverage labels (**SpEagle⁺**)
  - improves detection performance
- Computationally light version : SpLite (**SpLite⁺**)
  - significant speed-up

Thank You!

Stony Brook University
Computer Science

Code and Data available:

http://shebuti.com/collective-opinion-spam-detection/

srayana@cs.stonybrook.edu

http://www.cs.stonybrook.edu/~datalab/