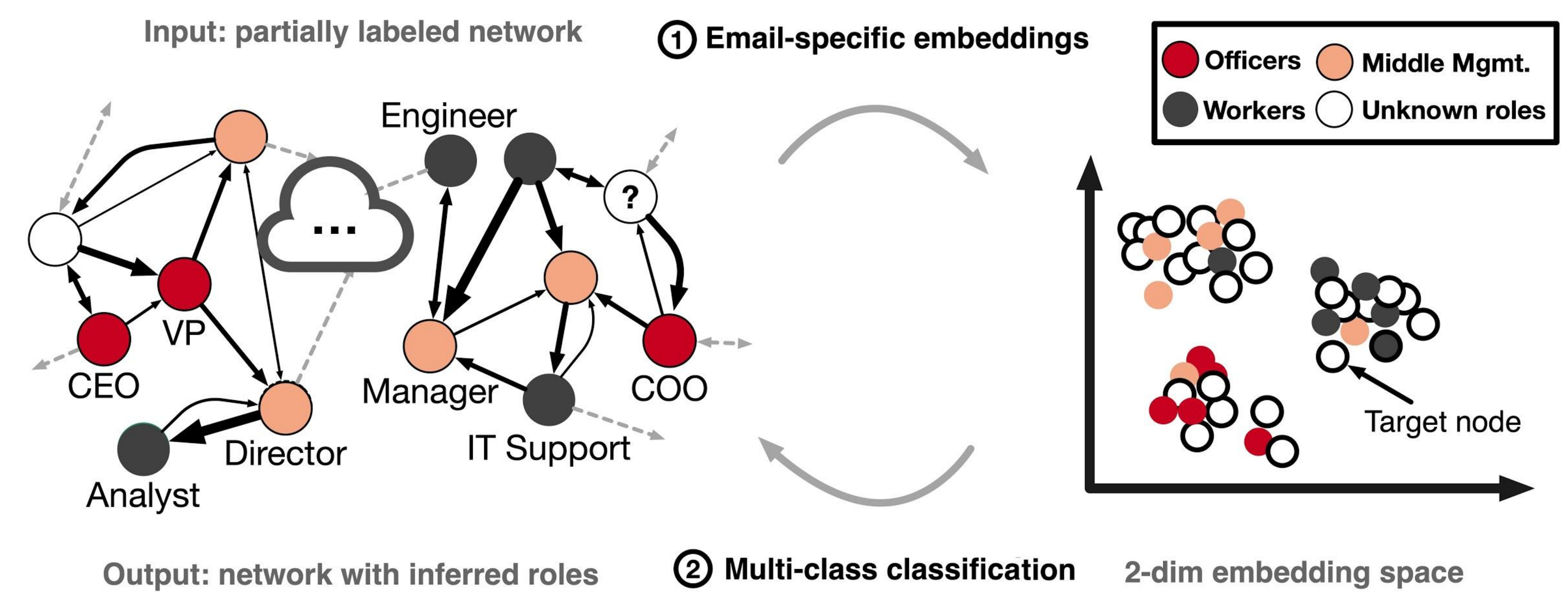
Smart Roles: Inferring Professional Roles in Email Networks

Di Jin[1,*] Mark Heimann[1,*] Tara Safavi[1] Mengdi Wang[2] Wei Lee[3] Lindsay Snider[3] and Danai Koutra[1]
[1]University of Michigan, Ann Arbor [2]University of Pittsburgh [3]Trove AI

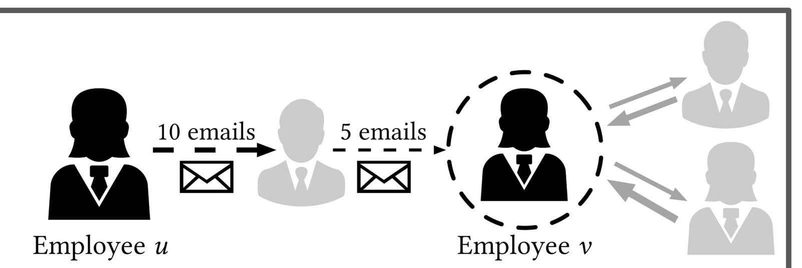
Motivation: bring order to flooded email inboxes
- *Prioritize* emails from important senders
- *Recommend* useful connections

Goal: Infer professional roles of email users

Intuition: Professional role inference ≈ *structural* role inference in email networks
Method: Design *node embedding* method, EMBER, to *efficiently* capture structural roles



S1: *Capture* each user's *local* structure



Email communication is...	We do...
Higher-order (indirect connections matter)	Define *structural behavior histograms* based on paths of communication up to length K $\mathbf{b}_u^+ = \sum_{k=0}^{K} \delta^k \, \mathbf{b}_u^{k+}$
Weighted by # of emails two users exchange	Make histograms bin *path weights* between a user and her connections $b_{u,d}^{k+} = \sum_{v \in D_u^{k+}} \texttt{path\_weight}\left(\mathcal{P}_{u \to v}^{k+}\right)$
Directed from sender to receiver	Create, concatenate separate histograms from paths following *outgoing* and *incoming* edges $\mathbf{b}_u = [\mathbf{b}_u^+, \mathbf{b}_u^-]$

S2: *Embed* users by comparing their local structure to a small number of *landmark users*

Embeddings $\tilde{\mathbf{Y}} = \mathbf{C}\mathbf{U}\boldsymbol{\Sigma}^{1/2}$

User-to-*landmark* similarities
- Sample landmarks ~ degree
- Structural similarity between users:

$$sim(u, v) = e^{-||\mathbf{b}_u - \mathbf{b}_v||}$$

Structural behavior histograms

From SVD of psuedoinverse of pairwise landmark similarities

- *Approximate* decomposition of pairwise user structural similarity matrix
- Can embed only important *subsets* of users

New email network datasets collected from Trove AI
- **Multiple** companies
- **Varying** company sizes

Group users' job titles into broad categories:
- **Officers**, middle **management**, and **workers**

Infer role of employees who do not provide titles

	Employees	Connections	Email exchanges	# Officers | Mid. mgmt | Workers
Trove-19	19	47	274	4 | 10 | 5
Trove-98	98	101	1769	53 | 32 | 13
Trove-141	141	1 242	9565	23 | 79 | 39
Trove-183	183	3136	21 655	16 | 133 | 34
Trove-318	318	1026	12 643	30 | 210 | 78
Trove-2K	2 414	16 281	183 443	495 | 1 300 | 620
Trove	9 989 507	40 290 044	568 678 419	495 | 1 300 | 620
Enron	75 416	319 935	2 064 442	31 | 44 | 41

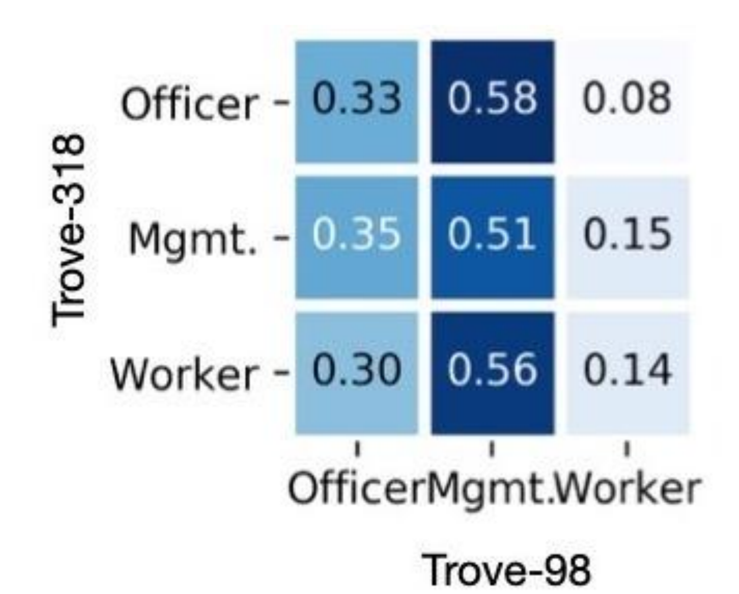
Effective: quantitative improvement in classification accuracy

	State-of-the-art baselines representing several families of techniques									Unweighted/undirected variants			
	SNA	RolX	LinBP	LINE	DeepWalk	node2vec	struc2vec	DNGR	Graphwave	EMBER-U	EMBER-D	EMBER-W	**EMBER**
Trove-318	.7605	.5670	.6908	.6618	.7602	.7648	.7799	.7131	.7685	.7749	.7563	.7625	**.8045***
Trove-183	.7648	.5787	.7718	.5657	.8071	.8223	**.8264**	4925	.6391	.7986	.7838	.8186	.8241
Trove-141	.6738	.5591	.7409	.7102	.7191	.7474	.7391	.6235	.7112	.7291	.7309	.6971	**.7568***
Trove-98	.6676	.5177	.6323	.6872	.5587	.6198	.6498	.5329	**.7177***	.6040	.5857	.6333	.6911
Trove-19	.5429	.6981	.6248	.7184	.5531	.5959	.6102	.6089	.7157	.6837	.7204	.6939	**.7337***
Trove-2K	.6305	.5212	.6622	.6771	.6769	.6780	**.6802**	.6527	.6594	.6689	.6345	.6677	.6745
Trove	.6633	.5280	5454	—	.6866	.6951	—	—	—	.6905	.7141	.7122	**.7162***
Enron	.6205	.5197	.5000	.6931	.7201	.7389	—	.5709	—	**.7393**	.7347	.7305	.7305

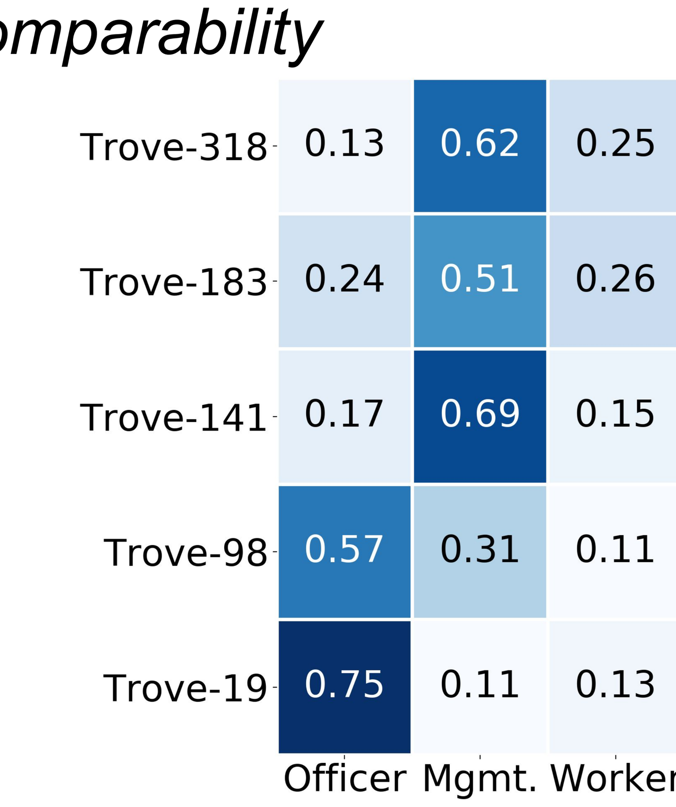
Practical: scales to *millions* of users

	Trove-318	Trove-2K	Trove	Enron
SNA	6.32	16.45	3193.26	333.33
RolX	**0.14**	**0.16**	2150.53	205.92
LinBP	0.54	2.88	14607.44	1038.09
LINE	171.95	153.12	>12h	267.48
DeepWalk	3.12	21.59	2464.13	255.84
node2vec	2.85	24.55	3484.05	254.60
struc2vec	17.48	188.65	>12h	29286.38
DNGR	21.05	72.83	>12h	>12h
Graphwave	2.73	5.66	>12h	>12h
EMBER	2.50	16.87	**830.80**	**84.98**

Insightful: can analyze role *comparability*



Mapping roles across different size companies



Mapping professors to professional roles

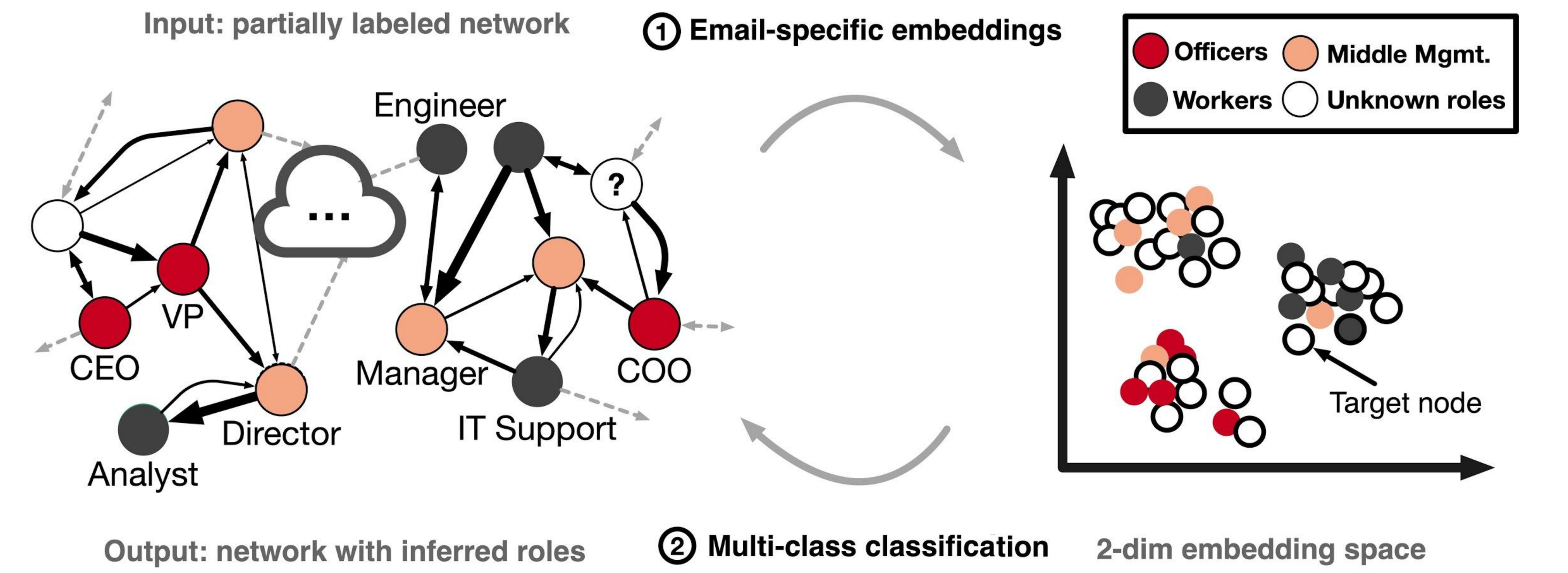
Smart Roles: Inferring Professional Roles in Email Networks

Di Jin[1]*, Mark Heimann[1]*, Tara Safavi[1], Mengdi Wang[2], Wei Lee[3], Lindsay Snider[3], and Danai Koutra[1]
[1]University of Michigan, Ann Arbor [2]University of Pittsburgh [3]Trove AI

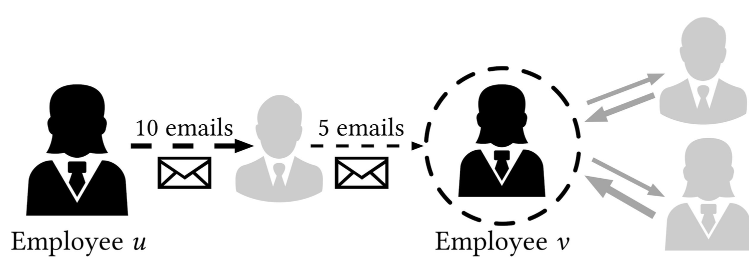
Motivation: bring order to flooded email inboxes
- *Prioritize* emails from important senders
- *Recommend* useful connections

Goal: Infer professional roles of email users

Intuition: Professional role inference ≈ *structural* role inference in email networks
Method: Design *node embedding* method, EMBER, to *efficiently* capture structural roles



S1: Capture each user's local structure



Email is:

Higher-order (indirect connections matter)
- Form *structural behavior histograms* based on patterns of length ≤K paths of communication

$$\mathbf{b}_u^+ = \sum_{k=0}^{K} \delta^k \, \mathbf{b}_u^{k+}$$

Weighted by # of emails two users exchange
- Bin path weights between a user and connections

$$b_{u,d}^{k+} = \sum_{v \in D_u^{k+}} \texttt{path\_weight}\left(\mathcal{P}_{u \to v}^{k+}\right)$$

Directed from sender to receiver
- Create. concatenate separate histograms from paths following outgoing and incoming edges

$$\mathbf{b}_u = [\mathbf{b}_u^+, \mathbf{b}_u^-]$$

S2: Embed local structure

$$\tilde{\mathrm{Y}} = \mathrm{C}\mathrm{U}\Sigma^{1/2}$$

User-to-*landmark* similarities
- Sample landmarks ~ degree
- Structural similarity between users:

From SVD of psuedoinverse of pairwise landmark similarities

$$sim(u, v) = e^{-||\mathbf{b}_u - \mathbf{b}_v||}$$

- *Approximate* decomposition of pairwise user structural similarity matrix
- Can embed only important *subsets* of users

New email network datasets collected from Trove AI
- **Multiple** companies
- **Varying** company sizes

Group users' job titles into broad categories:
- **Officers**, middle **management**, and **workers**

Infer role of employees who do not provide titles

	Employees	Connections	Email exchanges	# Officers | Mid. mgmt | Workers
Trove-19	19	47	274	4 | 10 | 5
Trove-98	98	101	1769	53 | 32 | 13
Trove-141	141	1 242	9565	23 | 79 | 39
Trove-183	183	3136	21 655	16 | 133 | 34
Trove-318	318	1026	12 643	30 | 210 | 78
Trove-2K	2 414	16 281	183 443	495 | 1 300 | 620
Trove	9 989 507	40 290 044	568 678 419	495 | 1 300 | 620
Enron	75 416	319 935	2 064 442	31 | 44 | 41

Effective: quantitative improvement in classification accuracy

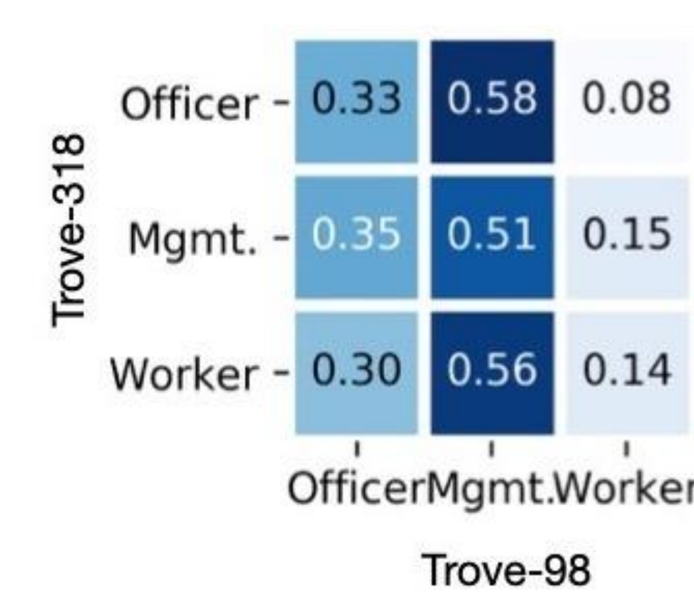
State-of-the-art baselines, several families of techniques | Unweighted/undirected variants

	SNA	RolX	LinBP	LINE	DeepWalk	node2vec	struc2vec	DNGR	Graphwave	EMBER-U	EMBER-D	EMBER-W	**EMBER**
Trove-318	.7605	.5670	.6908	.6618	.7602	.7648	.7799	.7131	.7685	.7749	.7563	.7625	**.8045***
Trove-183	.7648	.5787	.7718	.5657	.8071	.8223	**.8264**	4925	.6391	.7986	.7838	.8186	.8241
Trove-141	.6738	.5591	.7409	.7102	.7191	.7474	.7391	.6235	.7112	.7291	.7309	.6971	**.7568***
Trove-98	.6676	.5177	.6323	.6872	.5587	.6198	.6498	.5329	**.7177***	.6040	.5857	.6333	.6911
Trove-19	.5429	.6981	.6248	.7184	.5531	.5959	.6102	.6089	.7157	.6837	.7204	.6939	**.7337***
Trove-2K	.6305	.5212	.6622	.6771	.6769	.6780	**.6802**	.6527	.6594	.6689	.6345	.6677	.6745
Trove	.6633	.5280	5454	—	.6866	.6951	—	—	—	.6905	.7141	.7122	**.7162***
Enron	.6205	.5197	.5000	.6931	.7201	.7389	—	.5709	—	**.7393**	.7347	.7305	.7305

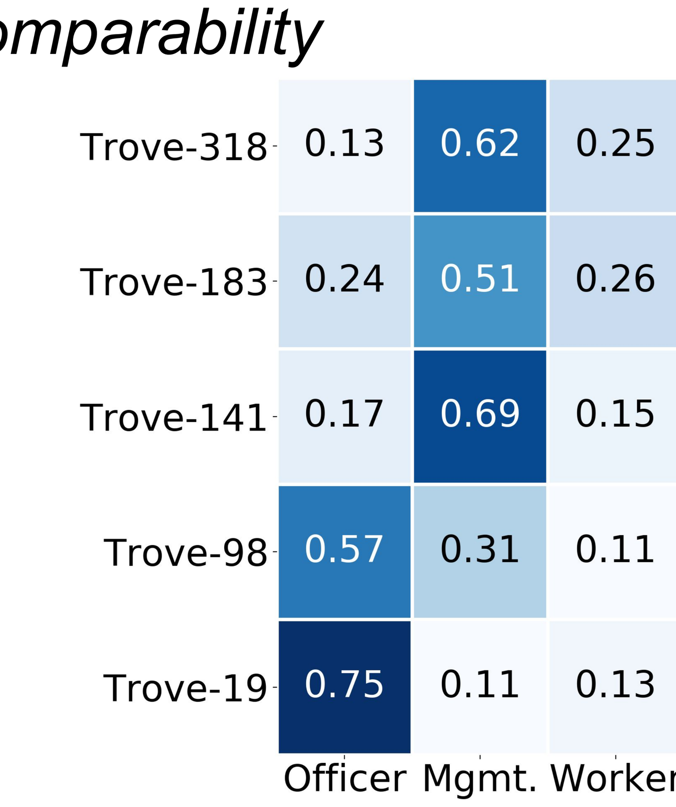
Practical: scales to *millions* of users

	Trove-318	Trove-2K	Trove	Enron
SNA	6.32	16.45	3193.26	333.33
RolX	**0.14**	**0.16**	2150.53	205.92
LinBP	0.54	2.88	14607.44	1038.09
LINE	171.95	153.12	>12h	267.48
DeepWalk	3.12	21.59	2464.13	255.84
node2vec	2.85	24.55	3484.05	254.60
struc2vec	17.48	188.65	>12h	29286.38
DNGR	21.05	72.83	>12h	>12h
Graphwave	2.73	5.66	>12h	>12h
EMBER	2.50	16.87	**830.80**	**84.98**

Insightful: can analyze role *comparability*



Mapping roles across different size companies



Mapping professors to professional roles