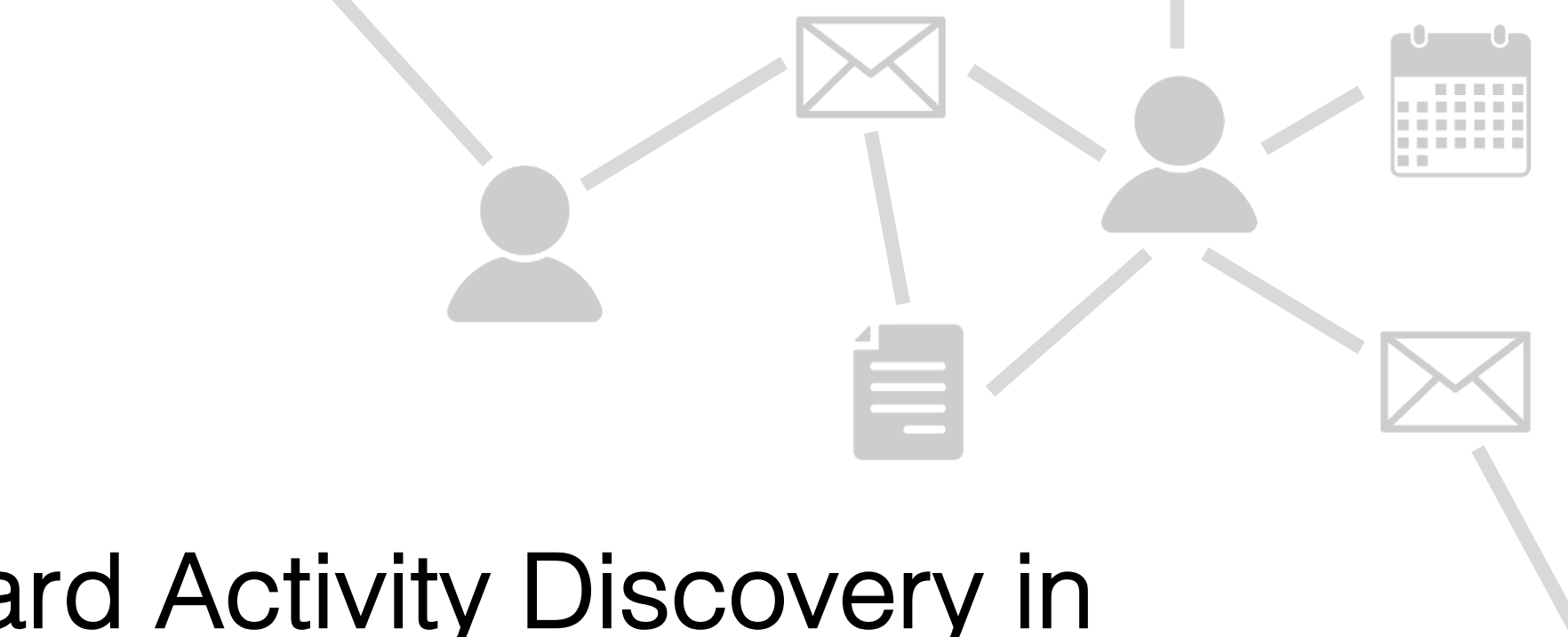


Toward Activity Discovery in the Personal Web

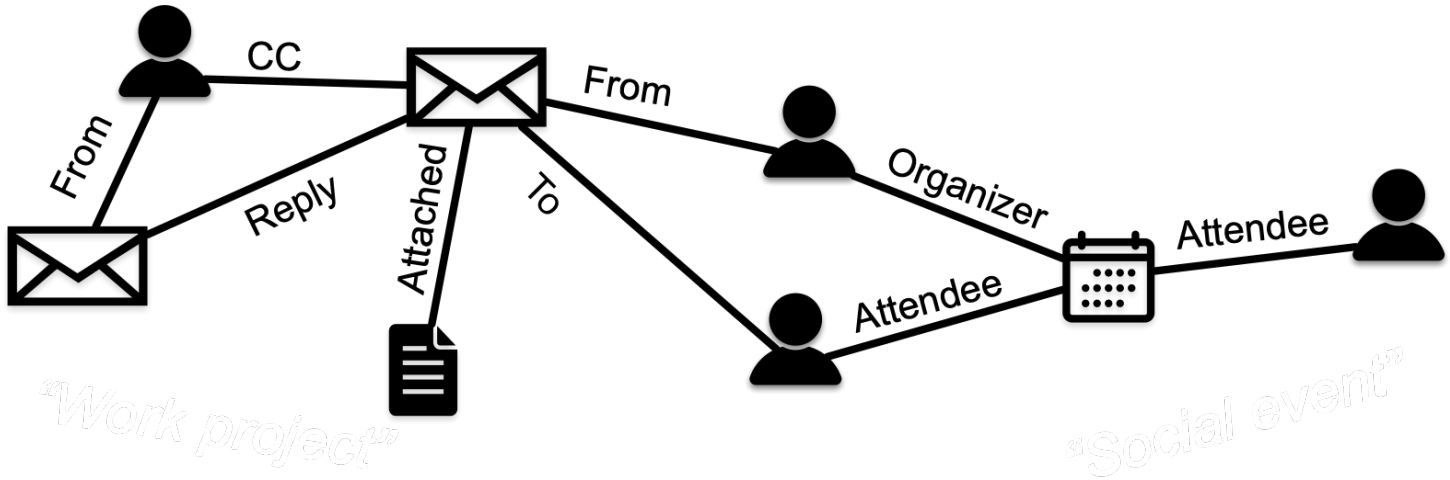
Tara Safavi (Michigan), Adam Fourney (MSR), Robert Sim (MSR),
Marcin Juraszek (MS), Shane Williams (MSR), Ned Friend (MS),
Danai Koutra (Michigan), Paul N. Bennett (MSR)

Activity discovery

Identify high-level activities from low-level entities in an individual's "personal web"

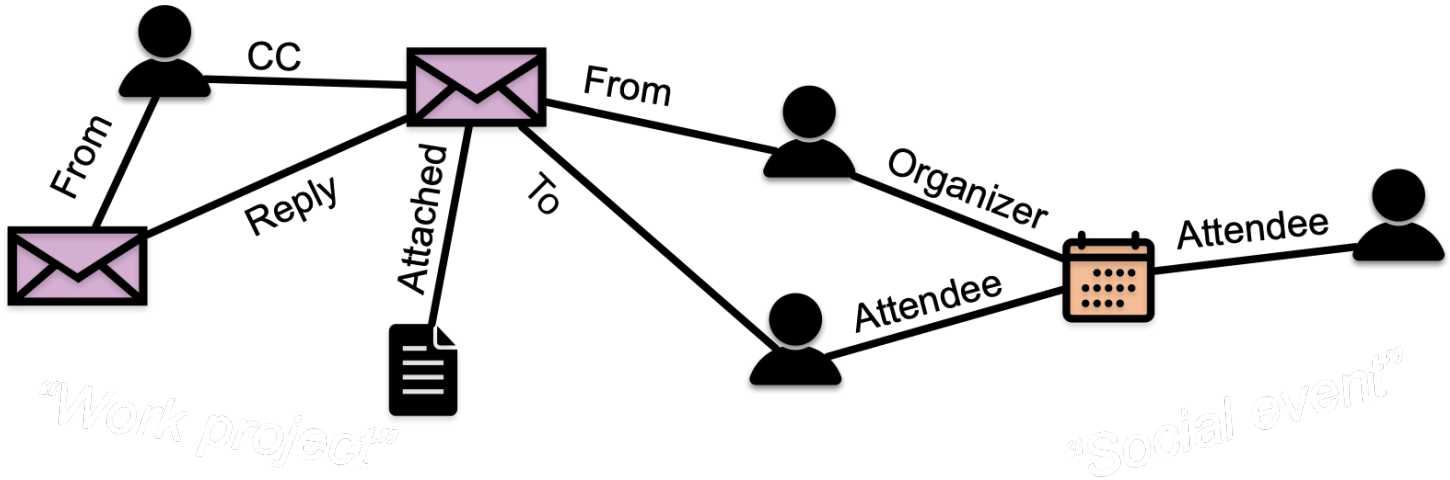
Activity discovery

Identify high-level activities from low-level entities in an individual's "personal web"



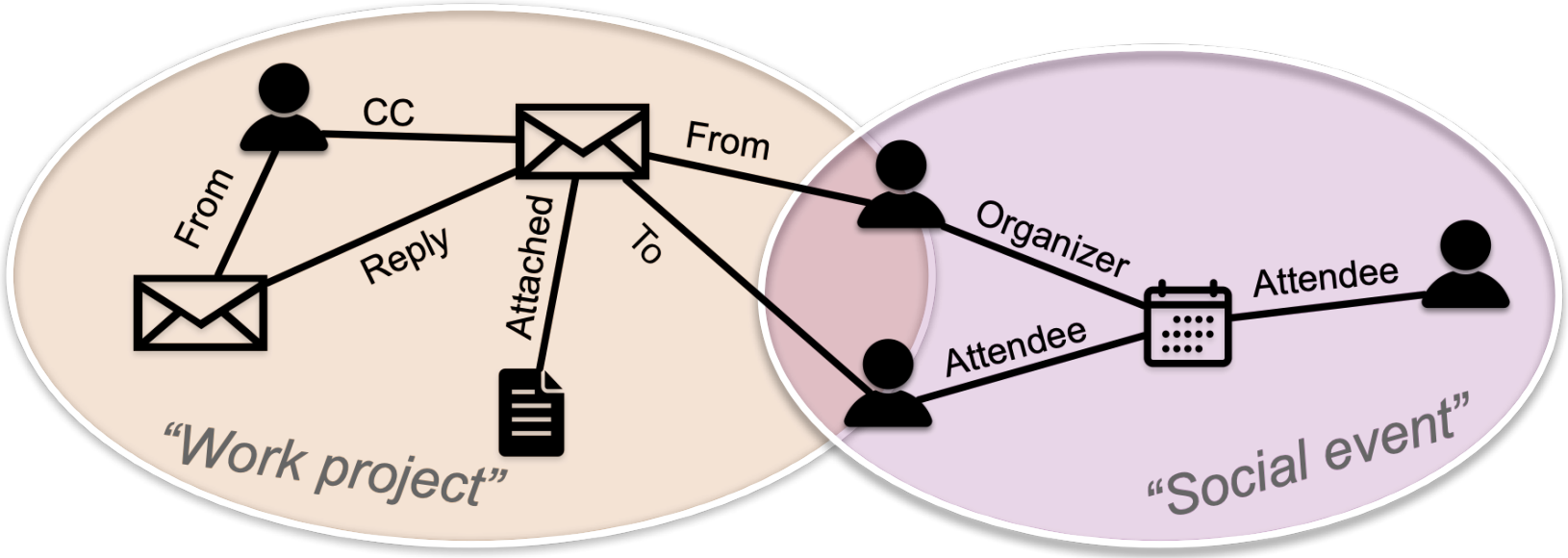
Activity discovery

Identify high-level activities from low-level entities in an individual's "personal web"



Activity discovery

Identify high-level activities from low-level entities in an individual's "personal web"



Projects, hobbies, tasks

Why activity discovery?

Task detection and reflection

Entity search and recommendation

Email prioritization and filtering

Why activity discovery?

Help individuals better organize, retrieve, and manage their own personal information

Related work

Personal information management

- [Dumais+ SIGIR03] [Jones ARIST07] [Bendersky+ WSDM17] [Wang+ WSDM18]

Activity modeling

- [Dredze+ IUI06] [Shen+ IUI06] [Qadir+ NAACL16]

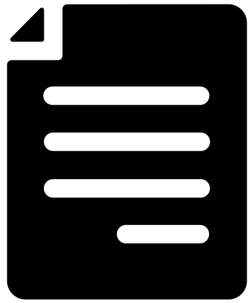
Graph-based similarity search

- [McAuley and Leskovec NeurIPS13] [Grover and Leskovec KDD16] [Jin+ KDD19]

Problem statement

Input: Entity **e** from heterogeneous personal information collection

Project X proposal

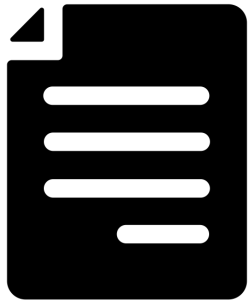


Problem statement

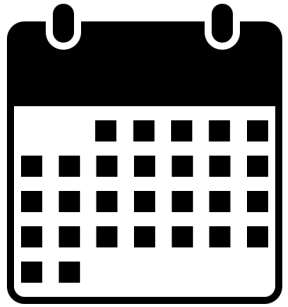
Input: Entity **e** from heterogeneous personal information collection

Output: "Similar" or "related" entities **e'** based on activities

Project X proposal



Meeting about Project X



Problem statement

Input: Entity **e** from heterogeneous personal information collection

Output: "Similar" or "related" entities **e'** based on activities

Requirements: Handle private, unlabeled, evolving data

Problem statement

Input: Entity **e** from heterogeneous personal information collection

Output: "Similar" or "related" entities **e'** based on activities

Requirements: Handle private, unlabeled, evolving data

Learn a model per individual on-device

Problem statement

Input: Entity **e** from heterogeneous personal information collection

Output: "Similar" or "related" entities **e'** based on activities

Requirements: Handle private, unlabeled, evolving data

Don't require labels, but incorporate them if available

Problem statement

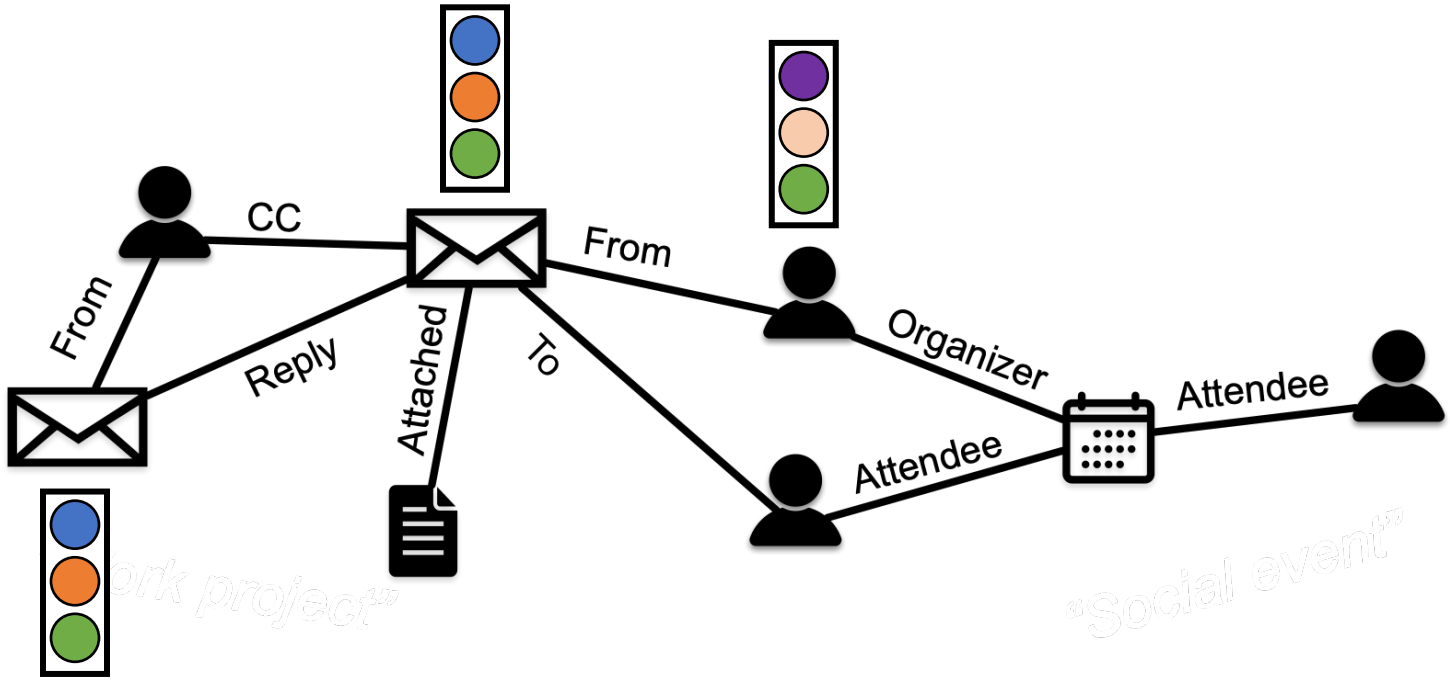
Input: Entity **e** from heterogeneous personal information collection

Output: "Similar" or "related" entities **e'** based on activities

Requirements: Handle private, unlabeled, evolving data

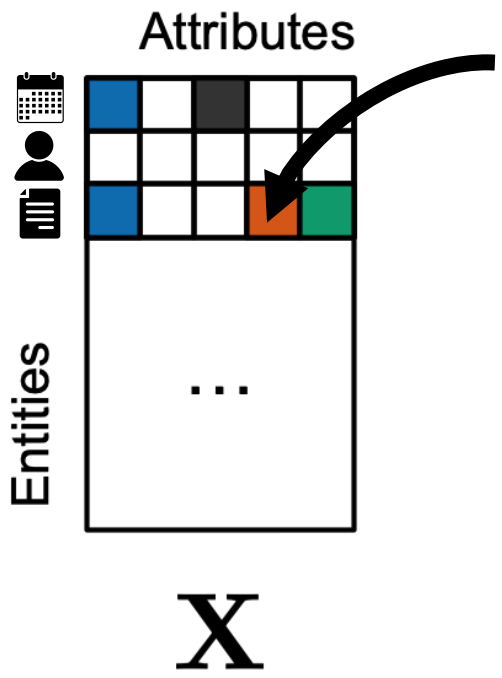
Incrementally update model as new data arrive

Intuition: Learn activity-based vector representations of entities in heterogeneous personal web



Graph-based representations

Seed matrix X: Map entities to "activity-related" attributes



Attributes: Noun-phrase frequencies, latent topic membership strengths, user labels, etc

"Entity i participates in attribute j with strength k"

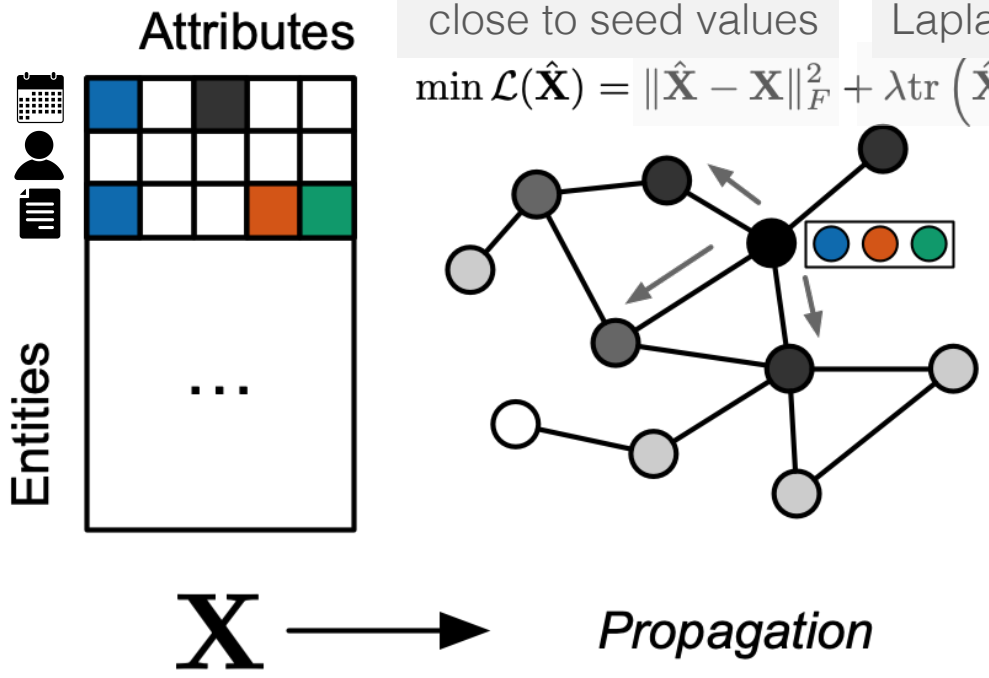
Graph-based representations

Propagate seed values over heterogeneous personal web

Constraint 1: Learned values close to seed values

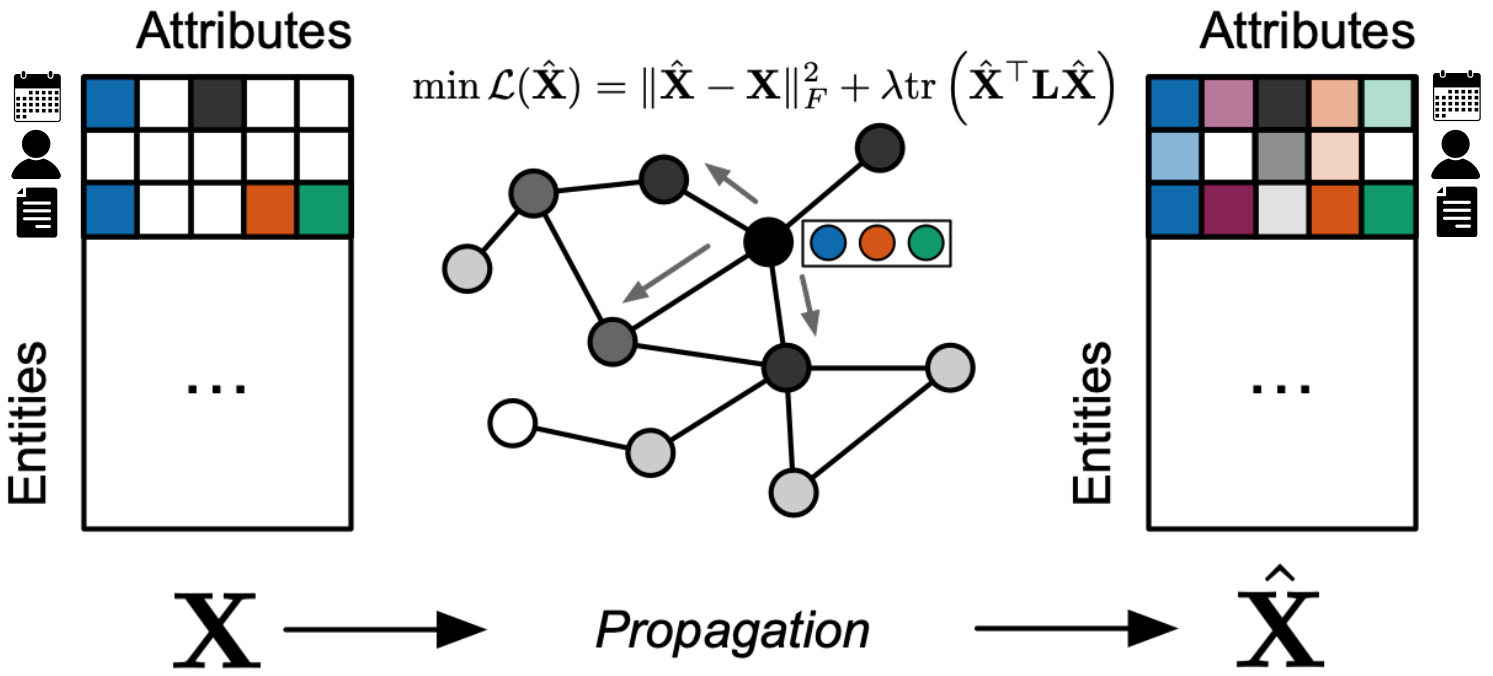
Constraint 2: Graph smoothness with Laplacian matrix L

$$\min \mathcal{L}(\hat{\mathbf{X}}) = \|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 + \lambda \mathrm{tr}\left(\hat{\mathbf{X}}^\top \mathbf{L} \hat{\mathbf{X}}\right)$$



Graph-based representations

Representation for i^{th} entity is i^{th} row of propagation matrix



Represent all heterogeneous entities in the same space

Model inference

All data given up front, learn model once

O(mp) for m edges and p attributes with Jacobi iteration

$$\hat{\mathbf{X}} = (\mathbf{I} + \mathbf{L})^{-1}\mathbf{X}$$

Offline

$$\min \mathcal{L}(\hat{\mathbf{X}}) = \|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 + \lambda \mathrm{tr}\left(\hat{\mathbf{X}}^\top \mathbf{L}\hat{\mathbf{X}}\right)$$

Model inference

All data given up front, learn model once

O(mp) for m edges and p attributes with Jacobi iteration

$$\hat{\mathbf{X}} = (\mathbf{I} + \mathbf{L})^{-1}\mathbf{X}$$

Offline

Derive exact online updates of representations via outer product

- $\mathbf{u}$: Update strengths for each entity
- $\mathbf{v}$: Update values from each attribute

Linear in max{# of edges, # of new attributes}

$$\begin{aligned}\Delta\hat{\mathbf{X}} &= \hat{\mathbf{X}}_{\text{new}} - \hat{\mathbf{X}} \\ &= \mathbf{u}\mathbf{v}^{\top}\end{aligned}$$

Online

$$\min \mathcal{L}(\hat{\mathbf{X}}) = \|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 + \lambda \text{tr}\left(\hat{\mathbf{X}}^{\top}\mathbf{L}\hat{\mathbf{X}}\right)$$

Intrinsic evaluation: Consenting participants conduct evaluations over their own data in a privacy-preserving evaluation setup

Example task

Entity A: document

https://en.wikipedia.org/wiki/Peregrine_falcon

Title: Peregrine falcon - Wikipedia

Last access: 2019-06-10 10:40:52 AM

Entity B: email

Birdwatching photos from hiking trip

Timestamp: 2019-06-10 10:34:23 AM

From: teammate@company.com

To: me@company.com

Here are the bird photos from the hiking trip. Note the falcons...

Example task

Entity A: document

https://en.wikipedia.org/wiki/Peregrine_falcon
Title: Peregrine falcon - Wikipedia
Last access: 2019-06-10 10:40:52 AM

Entity B: email

Birdwatching photos from hiking trip
Timestamp: 2019-06-10 10:34:23 AM
From: teammate@company.com
To: me@company.com

Here are the bird photos from the hiking trip. Note the falcons...

Q1: Why do you think the system found these entities to be related?

Select all that apply, if any.

- ☐ **Not related**: The system is wrong. I cannot find any relationship between these entities
- ☐ **Unsure**: The system may have its reasons, but I don't recognize one or more of these entities
- ☐ **Low-level**: These entities correspond to the same short-term task, appointment, or goal (e.g., a meeting, a TODO)
- ☐ **Mid-level**: These entities correspond to the same long-term project or activity (e.g., a research project, a home remodel)
- ☐ **High-level**: These entities correspond to the same general life category, not necessarily with defined start or end dates (e.g., "Personal", "Professional", "School")
- ☐ **Other**: These entities are related for reasons not listed above

Q1: Scope of entity relatedness

Example task

Entity A: document

https://en.wikipedia.org/wiki/Peregrine_falcon

Title: Peregrine falcon - Wikipedia

Last access: 2019-06-10 10:40:52 AM

Entity B: email

Birdwatching photos from hiking trip

Timestamp: 2019-06-10 10:34:23 AM

From: teammate@company.com

To: me@company.com

Here are the bird photos from the hiking trip. Note the falcons...

Q2: In your opinion, how related are these entities?

- **Strongly related**: There is a clear, strong connection between these entities
- **Related**: There is a connection between these entities
- **Somewhat related**: I can see some connection between these entities
- **A little related**: There is a tenuous connection between these entities
- **Not related at all**: I cannot find any connection between these entities
- **Unsure**: I don't recognize one or more of these entities

Submit answers

Q2: Grade of entity relatedness

Aggregate performance

	Recall	Prec@1	Prec@2	Prec@3	Prec@4	Prec@5
	***A little related, somewhat related, related,* and *strongly related* pairs**					
People Overlap						
node2vec						
Ours-NP						
Ours-LSA						
	***Strongly related* pairs only**					
	Recall	Prec@1	Prec@2	Prec@3	Prec@4	Prec@5
People Overlap						
node2vec						
Ours-NP						
Ours-LSA						

Baselines: Email features [Dredze+ IUI06] and node2vec [Grover and Leskovec KDD16]

Our representations: Using noun phrases (NP) and latent topics (LSA) as attributes

Aggregate performance

	Recall	Prec@1	Prec@2	Prec@3	Prec@4	Prec@5
***A little related, somewhat related, related,* and *strongly related* pairs**						
People Overlap	0.450 ± 0.11	0.933 ± 0.13	0.933 ± 0.11	0.922 ± 0.11	0.933 ± 0.10	0.933 ± 0.11
node2vec	0.440 ± 0.10	0.900 ± 0.21	0.867 ± 0.16	0.844 ± 0.17	0.858 ± 0.14	0.867 ± 0.14
Ours-NP	0.444 ± 0.07	0.967 ± 0.10	0.933 ± 0.11	0.911 ± 0.12	0.900 ± 0.11	0.867 ± 0.13
Ours-LSA	0.478 ± 0.07	1.000 ± 0.00	0.983 ± 0.05	0.944 ± 0.10	0.925 ± 0.10	0.907 ± 0.12
***Strongly related* pairs only**						
	Recall	Prec@1	Prec@2	Prec@3	Prec@4	Prec@5
People Overlap	0.319 ± 0.11	0.370 ± 0.25	0.370 ± 0.20	0.290 ± 0.17	0.265 ± 0.14	0.247 ± 0.15
node2vec	0.447 ± 0.25	0.333 ± 0.31	0.352 ± 0.32	0.333 ± 0.27	0.306 ± 0.19	0.274 ± 0.16
Ours-NP	0.507 ± 0.24	0.519 ± 0.28	0.481 ± 0.21	0.420 ± 0.21	0.398 ± 0.21	0.356 ± 0.20
Ours-LSA	0.522 ± 0.23	0.407 ± 0.26	0.407 ± 0.19	0.383 ± 0.20	0.380 ± 0.21	0.356 ± 0.18

Our representations best at identifying items with strong relationships

Per-participant performance

	All pairs of entities										
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**Avg. grade (rank)**
People Overlap											
node2vec											
Ours-NP											
Ours-LSA											
	***Email-Email* pairs only**										
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**Avg. grade (rank)**
People Overlap											
node2vec											
Ours-NP											
Ours-LSA											

Per-participant performance

All pairs of entities											
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**Avg. grade (rank)**
People Overlap	2.00 (4)	**2.47 (1)**	2.67 (4)	1.87 (4)	2.13 (3)	**2.77 (1)**	**2.00 (1)**	2.00 (2)	2.00 (3)	2.43 (3)	2.22 ± 1.23 (2.60)
node2vec	**2.33 (1)**	2.40 (2)	3.07 (3)	1.93 (3)	1.73 (4)	2.33 (2)	1.87 (3)	1.80 (3)	1.93 (4)	2.20 (4)	2.16 ± 1.38 (2.90)
Ours-NP	2.27 (2)	1.93 (4)	**3.53 (1)**	**2.13 (1)**	2.60 (2)	2.27 (3)	1.87 (2)	1.80 (3)	**2.53 (1)**	**2.73 (1)**	2.37 ± 1.43 (**2.00**)
Ours-LSA	2.13 (3)	2.13 (3)	3.27 (2)	2.07 (2)	**2.80 (1)**	2.27 (3)	1.87 (2)	**2.27 (1)**	2.47 (2)	2.53 (2)	2.38 ± 1.38 (2.10)
***Email-Email* pairs only**											
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**Avg. grade (rank)**
People Overlap	2.60 (2)	**2.67 (1)**	2.44 (4)	1.75 (3)	2.13 (3)	2.55 (4)	1.69 (4)	**2.20 (1)**	2.33 (3)	2.46 (2)	2.26 ± 1.30 (2.70)
node2vec	2.60 (2)	1.88 (3)	2.78 (3)	1.80 (2)	1.73 (4)	**3.71 (1)**	**2.00 (1)**	1.00 (3)	1.62 (4)	2.14 (4)	2.07 ± 1.39 (2.70)
Ours-NP	2.40 (4)	1.83 (4)	**3.29 (1)**	1.67 (4)	2.33 (2)	3.62 (2)	**2.00 (1)**	1.00 (3)	**2.57 (1)**	**2.50 (1)**	2.40 ± 1.40 (2.30)
Ours-LSA	**2.80 (1)**	2.29 (2)	2.88 (2)	**2.00 (1)**	**2.79 (1)**	3.62 (2)	1.89 (3)	2.11 (2)	2.43 (2)	2.42 (3)	2.54 ± 1.30 (**1.90**)

Latent-topic representations find highly related pairs of emails

Extrinsic evaluation: Use email recipient recommendation on public dataset as proxy task

Recipient recommendation task

Avocado email dataset: 128 inboxes, around 500-12000 nodes

Test emails with 2+ recipients, remove last recipient on To line

Each method returns ranked list of candidate recipients

Recipient recommendation task

Baselines: Random ranking, email features [Qadir+ NAACL16], Euclidean distance with node2vec

Our representations: Euclidean distance with noun phrases (NP), topics (LSA + LDA) as attributes

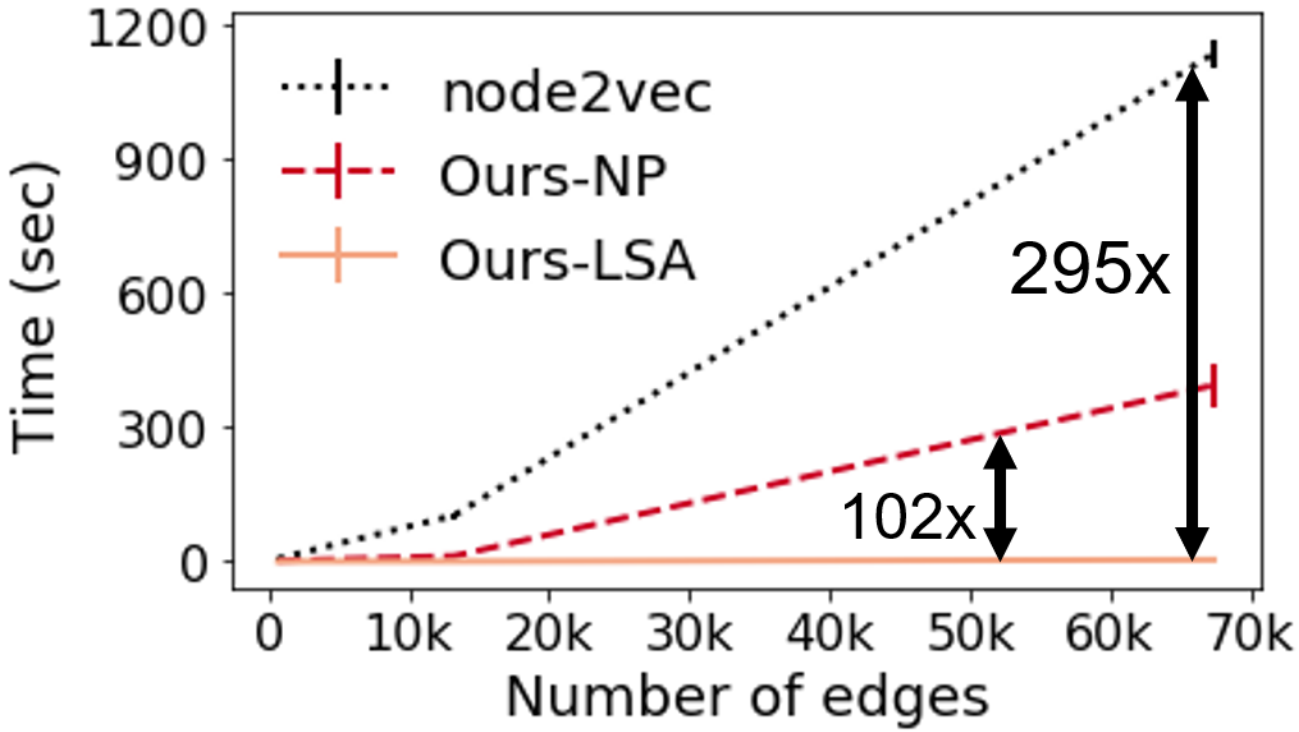
	Hits@1	Hits@2	MRR
Random			
Freq. Recipients			
Cond. On Sender			
Average NP			
node2vec			
Ours-NP, $\lambda = 10^{-1}$			
Ours-NP, $\lambda = 10^{0}$			
Ours-NP, $\lambda = 10^{2}$			
Ours-LSA			
Ours-LDA			

Recipient recommendation task

Our representations tie or outperform strong baselines, suggesting their versatility

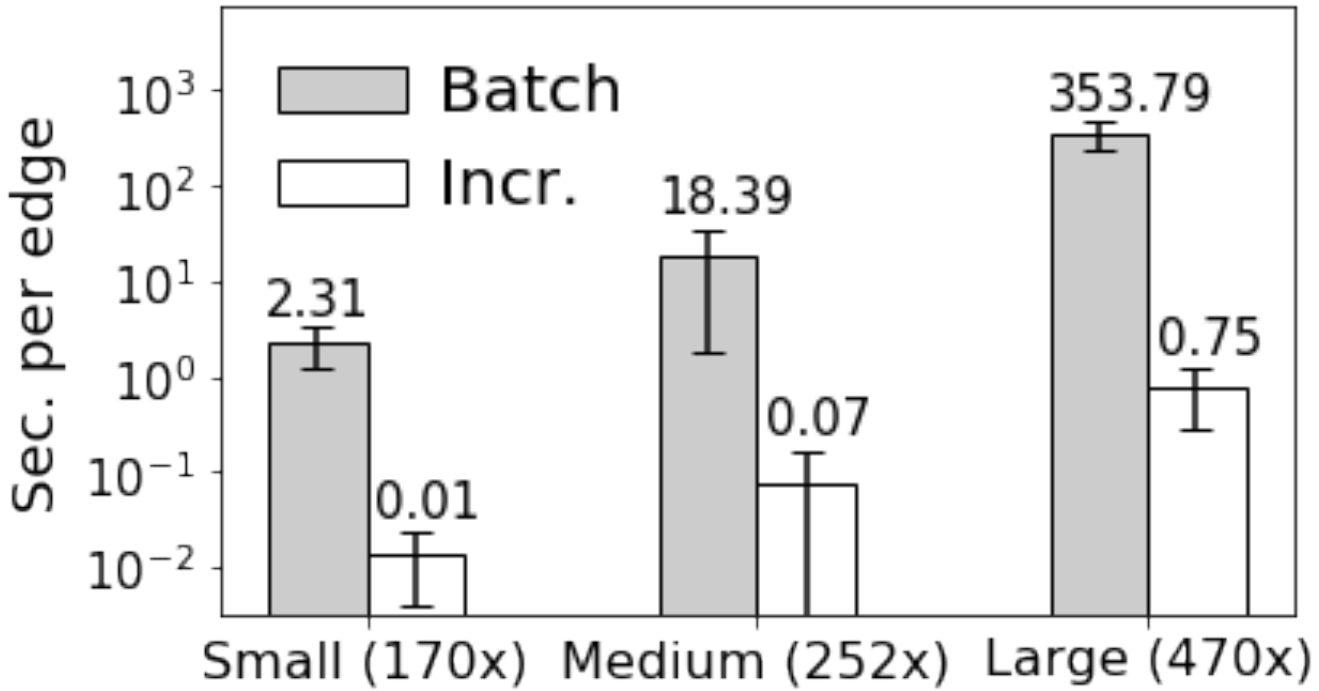
	Hits@1	Hits@2	MRR
Random	0.019 ± 0.023	0.038 ± 0.040	0.081 ± 0.060
Freq. Recipients	0.107 ± 0.106	0.184 ± 0.136	0.229 ± 0.105
Cond. On Sender	$0.143 \pm 0.094^{\dagger}$	$0.247 \pm 0.113^{\blacktriangle}$	$0.282 \pm 0.090^{\dagger}$
Average NP	0.128 ± 0.088	0.209 ± 0.119	0.259 ± 0.102
node2vec	0.062 ± 0.072	0.092 ± 0.108	0.126 ± 0.114
Ours-NP, $\lambda = 10^{-1}$	0.111 ± 0.059	0.182 ± 0.096	0.225 ± 0.082
Ours-NP, $\lambda = 10^{0}$	$0.158 \pm 0.084^{\blacktriangle}$	$0.247 \pm 0.105^{\blacktriangle}$	$0.290 \pm 0.089^{\blacktriangle}$
Ours-NP, $\lambda = 10^{2}$	$0.143 \pm 0.085^{\dagger}$	$0.225 \pm 0.112^{\dagger}$	$0.267 \pm 0.093^{\dagger}$
Ours-LSA	0.110 ± 0.093	0.180 ± 0.126	0.224 ± 0.111
Ours-LDA	0.082 ± 0.080	0.141 ± 0.123	0.189 ± 0.111

Offline scalability



Offline version up to 295x faster than traditional graph embedding

Online scalability



Online version up to 470x faster per edge than offline version

Summary

Activity discovery for personal information management

Graph-based representations learned offline + online

Extensive evaluation, both intrinsic and extrinsic

*Additional results in the paper + poster + supplemental slides

Graph-based representations for activity discovery in the personal web

THANK YOU + QUESTIONS

Tara Safavi

tsafavi@umich.edu

@tararootcake

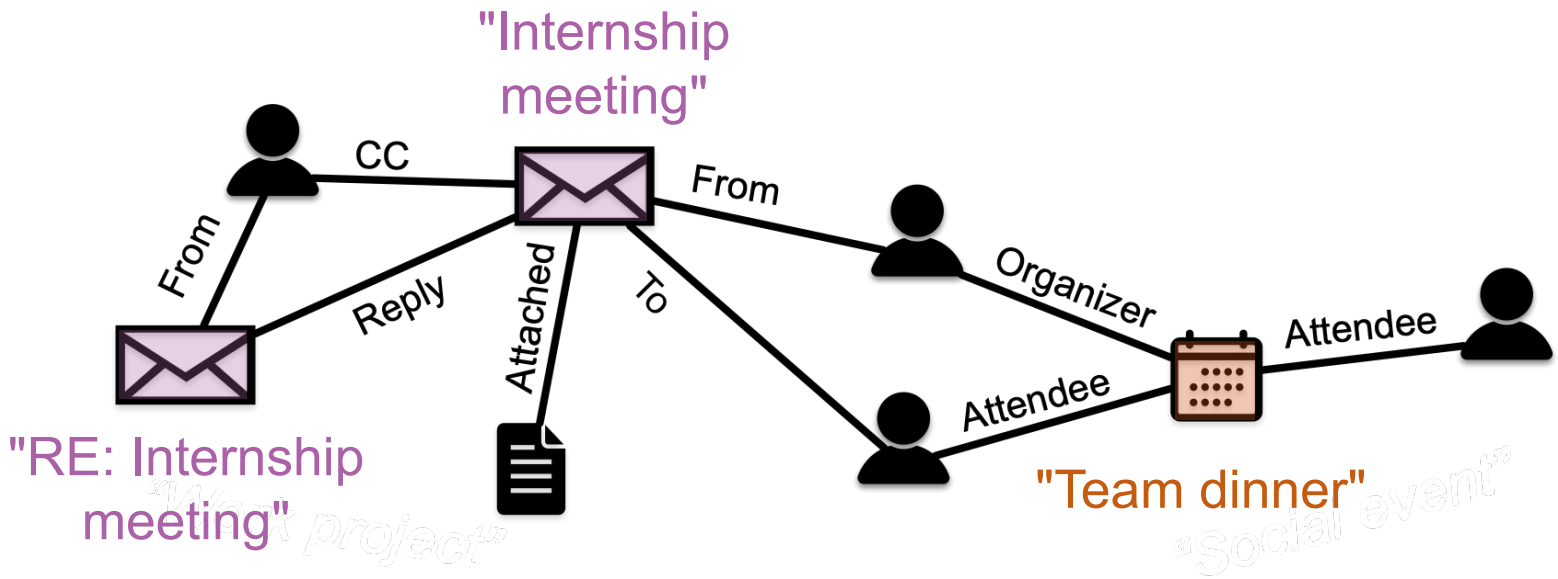
tsafavi.github.io

COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

Microsoft

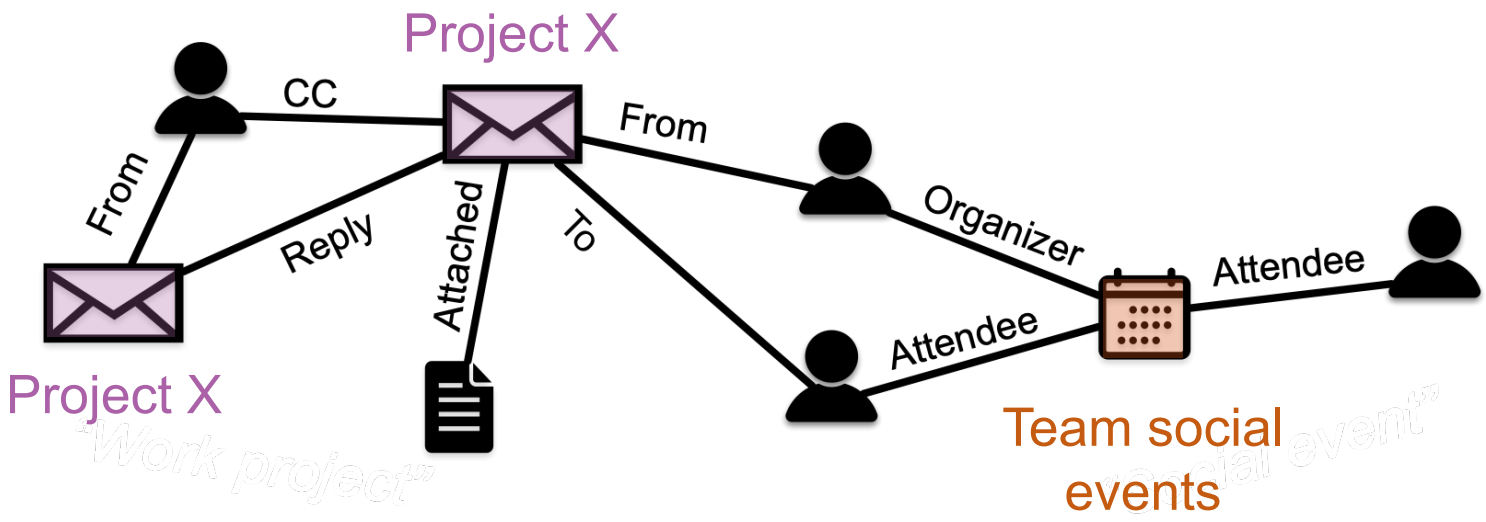
Activity-specific attributes

Unsupervised: Noun phrases, topics from text-bearing entities

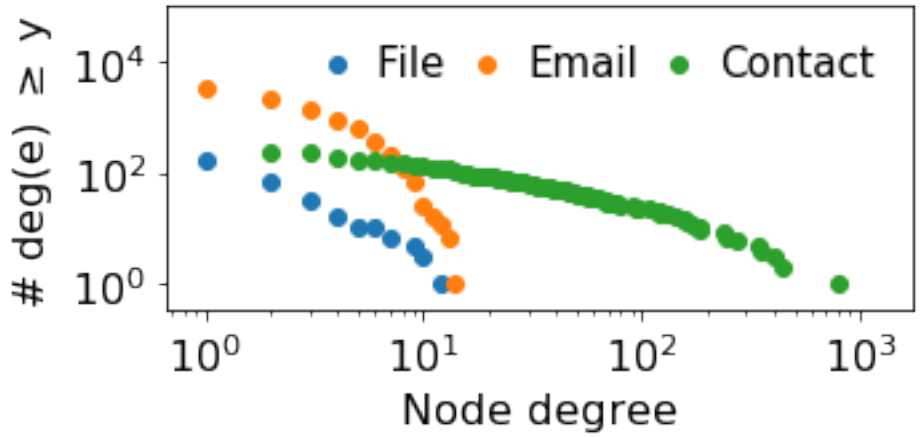


Activity-specific attributes

Semi-supervised: Entities have user-given labels

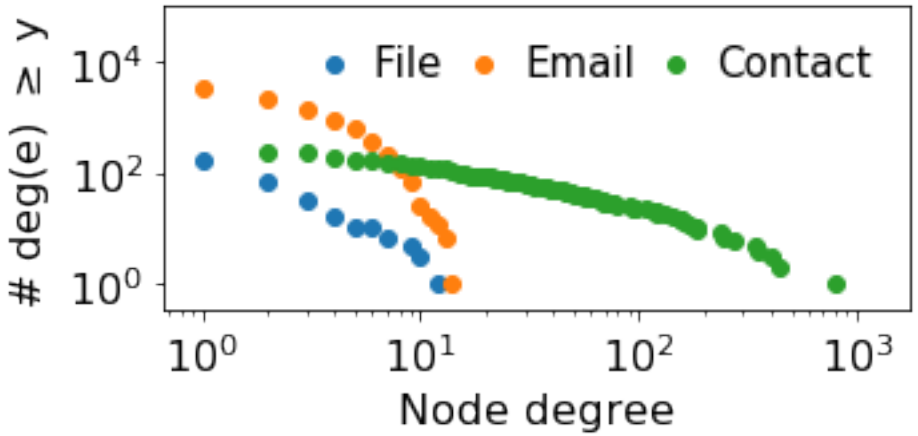


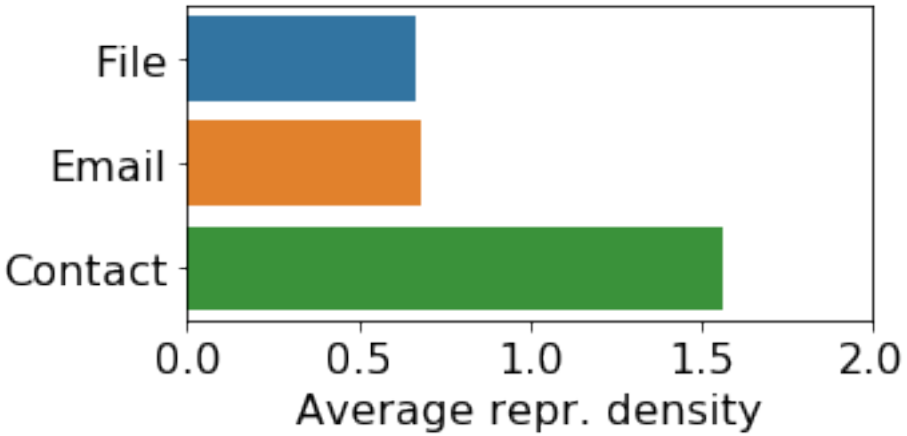
Implicitly capturing entity types



Graph construction leads to skewed degree distribution

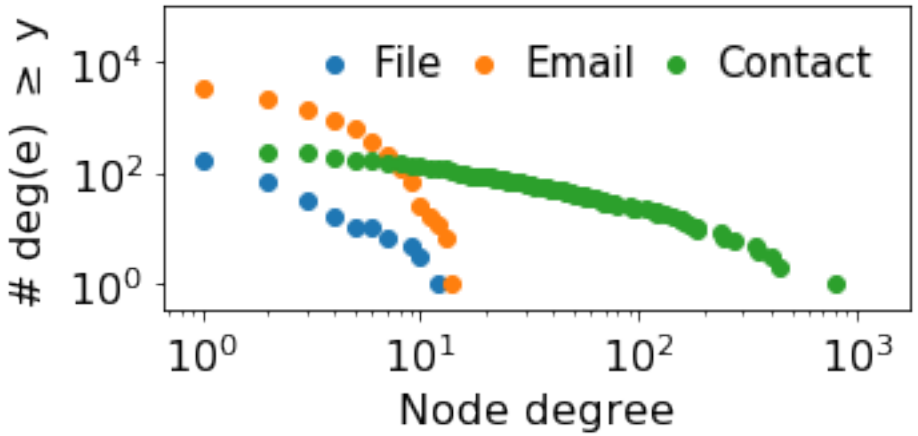
Implicitly capturing entity types

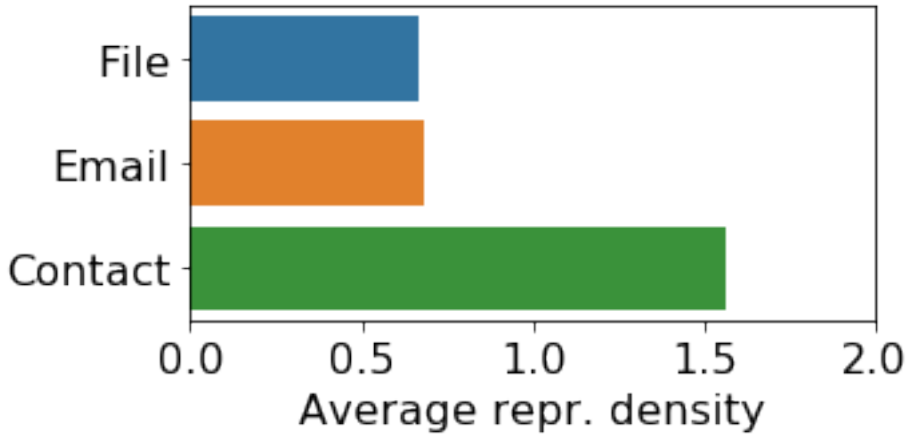




Propagation leads to higher density representations for entities with higher degree

Implicitly capturing entity types

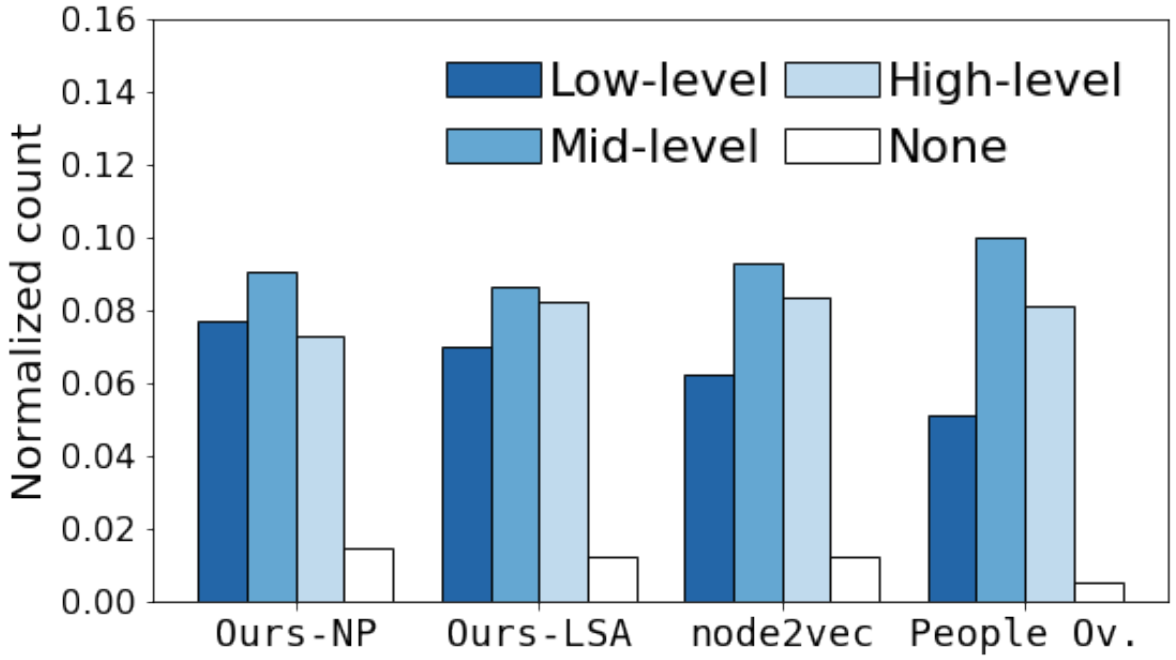




Intuition: People participate in more activities than emails, files

"Why do you think the system related this pair of entities?"

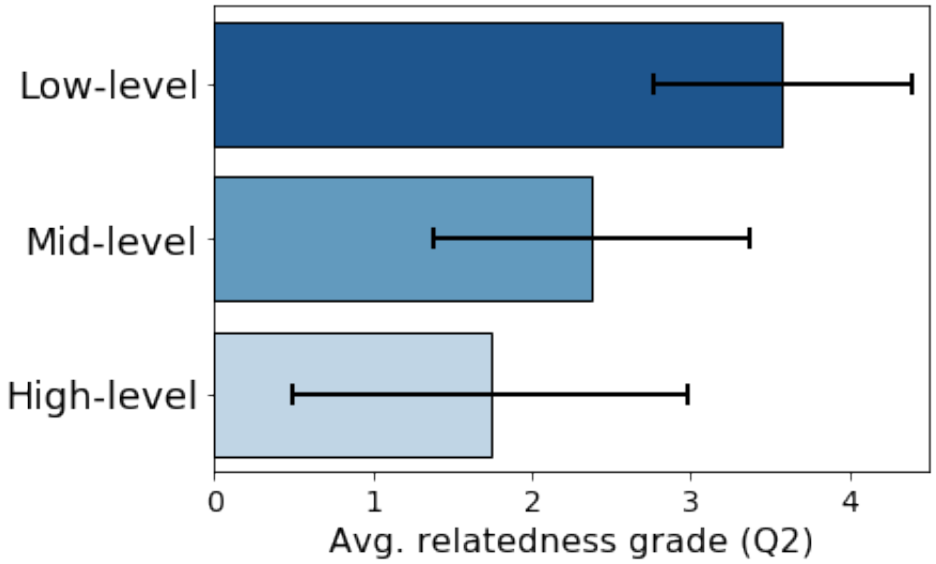
- ❑ **Low-level**: Short-term tasks or goals
- ❑ **Mid-level**: Long-term project or activity
- ❑ **High-level**: Same general life category



Our representations capture the most "low-level" relationships

"In your opinion, how related is this pair of entities?"

- Not related at all
- A little related
- Somewhat related
- Related
- Strongly related



Users tend to see pairs of items with "low-level" relationships as "strongly related"