Toward Activity Discovery in the Personal Web

Tara Safavi, Adam Fourney, Robert Sim, Marcin Juraszek, Shane Williams, Ned Friend, Danai Koutra, Paul N. Bennett

COMPUTER SCIENCE & ENGINEERING, UNIVERSITY OF MICHIGAN

Microsoft

Overview 1

Research problem

Identify **high-level "activities"** from low-level entities in individuals' heterogeneous personal information collections (**personal webs**) in a private, unsupervised, online manner.

Why activity discovery?

- Task detection and reflection
- Entity search and recommendation
- Email prioritization and filtering

Graph-based representations 2

INTUITION

- Model collection of personal information entities (emails, files, contacts, appts, web searches) as a **graph**
- Learn representations via **graph propagation** from a set of seeds

CONTRIBUTIONS

- Derive exact online updates of representations via outer product
- $\Delta\hat{\mathbf{X}}$: Update to entity representations
- $\mathbf{u}$: Update strengths for each entity
- $\mathbf{v}$: Update values from each attribute

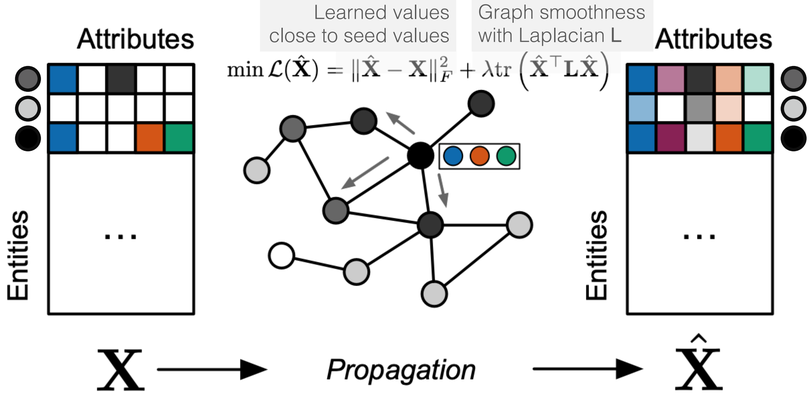
$$\Delta\hat{\mathbf{X}} = \hat{\mathbf{X}}_{\text{new}} - \hat{\mathbf{X}} = \mathbf{u}\mathbf{v}^\top$$

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

Constraint 1: Learned values close to seed values

Constraint 2: Graph smoothness with Laplacian L

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



Seeds: Noun phrases, latent topics, user labels indicating activities

Entity representations: Rows of matrix after propagation

References

[1] Dredze et al. Automatically classifying emails into activities. IUI 2006.
[2] Grover and Leskovec. node2vec: Scalable feature learning for networks. KDD 2016.
[3] Qadir et al. Activity modeling in email. NAACL-HLT 2016.

This material is partially based upon work supported by the National Science Foundation and an Army Young Investigator Award.

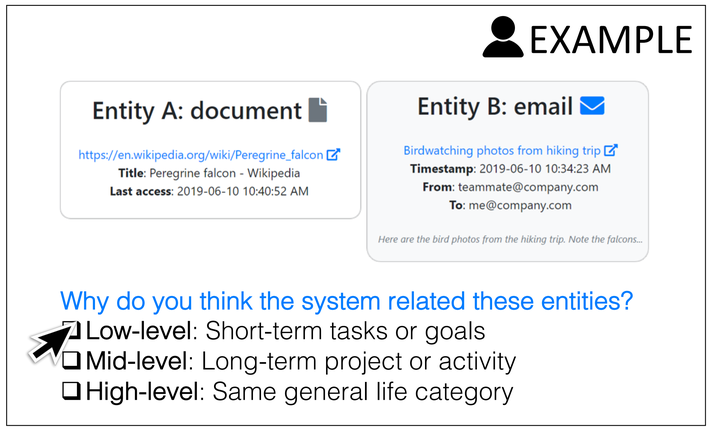
Intrinsic evaluation 3

Data

- Participants: 10 interns, researchers, managers
- 2-7 days of data from local logging application
- Recent emails, appts, contacts, searches, files
- Around 100 to 1k entities per participant
- Extract noun phrases (NP) and topics (LSA) from text

Privacy-preserving task setup

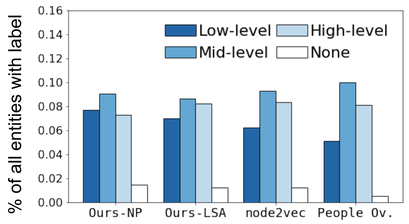
- Task hosted locally on participants' machines via USB
- Display pairs of personal information entities [1, 2]
- Participants rate the "activity relatedness" of pairs:
  - **Scope** (low-, mid-, high-level) and **grade** (0-4 points)
- All feedback anonymized and aggregated

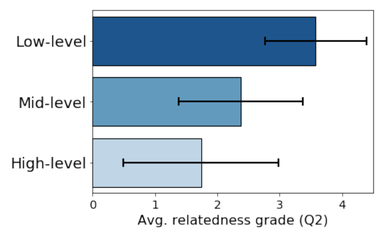


Participant # entities in G	P1 157	P2 258	P3 320	P4 303	P5 256	P6 291	P7 203	P8 232	P9 1468	P10 1637	Avg. grade (rank)
All pairs of entities											
People Overlap	2.00 (4)	2.47 (1)	2.67 (4)	1.87 (4)	2.77 (1)	**2.00 (1)**	2.00 (2)	2.00 (3)	2.43 (3)	2.13 (3)	2.22 ± 1.23 (2.60)
node2vec	2.33 (1)	2.40 (2)	3.07 (3)	1.93 (3)	2.33 (2)	1.87 (2)	1.80 (3)	1.93 (4)	2.20 (4)	1.73 (4)	2.16 ± 1.38 (2.80)
Ours-NP	2.27 (2)	1.93 (4)	**3.53 (1)**	**2.13 (1)**	2.27 (3)	1.87 (2)	1.80 (3)	**2.53 (1)**	**2.73 (1)**	2.60 (2)	2.37 ± 1.43 (2.00)
Ours-LSA	2.13 (3)	2.13 (3)	3.27 (2)	2.07 (2)	2.27 (3)	1.87 (2)	**2.27 (1)**	2.47 (2)	2.53 (2)	**2.80 (1)**	2.38 ± 1.38 (2.10)
***Email-Email* pairs only**											
People Overlap	2.60 (2)	**2.67 (1)**	2.44 (4)	1.75 (3)	2.55 (4)	1.69 (4)	**2.20 (1)**	2.33 (3)	2.46 (2)	2.13 (3)	2.26 ± 1.30 (2.70)
node2vec	2.60 (2)	1.88 (3)	2.78 (3)	1.80 (2)	**3.71 (1)**	**2.00 (1)**	1.00 (3)	1.62 (4)	2.14 (4)	1.73 (4)	2.07 ± 1.39 (2.70)
Ours-NP	2.40 (4)	1.83 (4)	**3.29 (1)**	1.67 (4)	3.62 (2)	**2.00 (1)**	1.00 (3)	**2.57 (1)**	**2.50 (1)**	2.33 (2)	2.40 ± 1.40 (2.30)
Ours-LSA	**2.80 (1)**	2.29 (2)	2.88 (2)	**2.00 (1)**	3.62 (2)	1.89 (3)	2.11 (2)	2.43 (2)	2.42 (3)	**2.79 (1)**	2.54 ± 1.30 (1.90)

The pairs retrieved by our representations were rated the most "activity-related" by participants, especially those in senior-level roles

Our representations perform best at identifying "low-level" relationships among entities: Short-term tasks and goals





Extrinsic evaluation 4

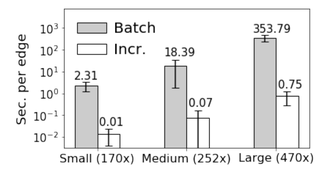
Data + task setup

- Avocado dataset: 128 inboxes, 500 to 12k entities per inbox
- Learn models on training emails, predict last recipient on test emails
- Baselines: Email features [3] and node2vec [2]

	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 ± 0.094†	0.247 ± 0.113▲	0.282 ± 0.090†
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 ± 0.084▲	0.247 ± 0.105▲	0.290 ± 0.089▲
~~Ours-NP, $\lambda = 10^{2}$~~	~~0.143 ± 0.085†~~	~~0.235 ± 0.112†~~	~~0.267 ± 0.093†~~
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

Our representations match or outperform strong baselines on the task, suggesting their versatility

Scalability 5



Online updates 470x faster than offline